HOME-Alcar: description of the CONLL format

Sergio Torres Aguilar[1], Dominique Stutzmann[1]

November 1, 2021

[1]Institut de Recherche et d'Histoire des Textes (CNRS)

Abstract

This document describes the CONLL format in the *HOME-Alcar* corpus of aligned and annotated cartularies, that is used to record the textual, linguistic, and graphic information. Each of the seventeen editions in the corpus is provided in a **.xlsx** column format, located and named as follows:

```
\Database\{CodeNameOfCartulary}\CONLL\{CodeNameOfCartulary\}_final_version_inner.xlsx
```

The resources in the corpus are so-called "cartularies", books in which are copied several legal texts, also called "acts" or "charters".

1 Col. A: INDEX

Sub-index for each charter: an integer from 1 to N. **Nota:** The index is reset when switching from one charter to the other. It is not an absolute index.

2 Col. B: WORD_x

Tokenized words according to the edited version of the charter.

3 Col. C: LEMMA

Lemma version (or dictionary version) for each token. We use a different set of labels according to the language. For French consult (`http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_principes_2.0.pdf`). For Latin consult (`http://www.glossaria.eu/sources/treetagger/classes.txt`)

4 Col. D: POS

Part-of-speech tagging: Grammatical tagging for each word. Lemma and POS are both automatically generated from the same lemmatizer.

5 Col. E: CASE

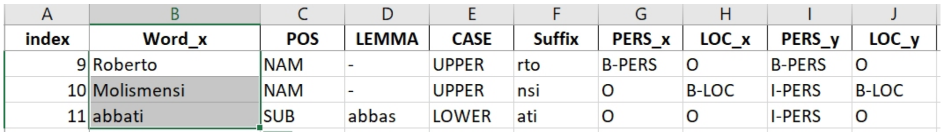
binary value according to the word cap : lowercase (LOWER) and uppercase (UPPER)

6 Col. F: Suffix

Affix placed after the stem of each word (normally three or two characters)

7 Col. G-H: PERS_x, and, optional, PERS_y

We use two annotation styles: the first one (named _x) includes only the proper names; the second one (named _y) is tagging the full entity including the proper name and co-occurrences as personal titles, dignities, functions, etc. (v.g. *presbiter, dominus, comes, miles*). In most cases the annotations are coincident since the entities do not always present personal collocations in diplomatics texts. The second style (_y) produces so-called "nested entities", when a name, esp. a place name, is included in the denomination of a person.



A	B	C	D	E	F	G	H	I	J
index	Word_x	POS	LEMMA	CASE	Suffix	PERS_x	LOC_x	PERS_y	LOC_y
9	Roberto	NAM	-	UPPER	rto	B-PERS	O	B-PERS	O
10	Molismensi	NAM	-	UPPER	nsi	O	B-LOC	I-PERS	B-LOC
11	abbati	SUB	abbas	LOWER	ati	O	O	I-PERS	O

Figure 1: In the example, the name “Roberto Molismensi abbati” (Robert, abbot of Molesme) contains a proper person name “Roberto” and a place name “Molismensi”

8 Col. H, or Col. I-J: LOC_x, and, optional, LOC_y

If there is a PERS_y annotation, then LOC_x is in column I and there is also a LOC_y annotation in column J. If there is no PERS_y annotation, then LOC_x is in column H and there are no LOC_y annotation.

The same annotation principle is used in the case of locations, but the differences are less evident between the two styles since in the case of LOC_Y annotation we do not tag the common words acting as co-occurrences (v.g. *territorium, ecclesia, terra, villa*, etc.)

9 Col. I or K: LANG

Language of the charters. Two values for French (FR) and Latin (LAT). Some french charters can contain some Latin formula and some Latin charters can

WORD	**PERS_x**	**LOC_x**	**PERS_y**	**LOC_y**
cum	O	O	O	O
Jacobus	B-PERS	O	B-PERS	O
presbiter	O	O	I-PERS	O
de	O	O	I-PERS	O
Petraficta	O	B-LOC	I-PERS	B-LOC
emisset	O	O	O	O

Table 1: Difference between the two annotation styles (X style). « Iacobus, presbyter of Petraficta » (Pierrefitte-sur-Seine)

WORD	**PERS_x**	**LOC_x**	**PERS_y**	**LOC_y**
Ego	O	O	O	O
Johannes	B-PERS	O	B-PERS	O
dictus	I-PERS	O	I-PERS	O
le	I-PERS	O	I-PERS	O
Bigot	I-PERS	O	I-PERS	O
miles	O	O	I-PERS	O
,	O	O	I-PERS	O
dominus	O	O	I-PERS	O
de	O	O	I-PERS	O
Condeto	O	B-LOC	I-PERS	B-LOC
supra	O	O	I-PERS	O
Rilum	O	B-LOC	I-PERS	B-LOC

Table 2: Difference between the two annotation styles (Y style). « I, Iohannes, called le Bigot, knight, lord of Condeto above Rilus (Condé-sur-Risle) »

contain french introduction or clauses. Values are done for the whole document, we do not provide sentence tagging.

10 Col. J or L: ACT

The « ACT » column contains the ID of each charter. IDs are generated following the natural numerical order from 1 to n. The naming pattern is {ACT_#} for the texts of the acts. Preliminary tables and other extraneous parts have other names, such as « Table ». When the original edition provided its own IDs, they are recorded in the column « Act_Original_ID » (see below).

11 Col. K or M: Word_y

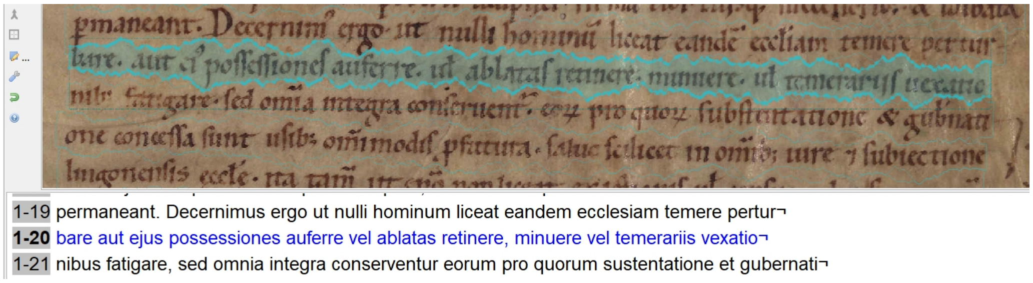
Tokenized words from the aligned text. The difference between Word_x (col. B) and Word_y is minimal and corresponds to the removal of foreign characters and some paratextual signs present in the original edition before loading the texts to Transkribus.

12 Col. L or N: LINES

Binary value. « regular » if the word is between the first and penultimate line position; « break » if the word is the last word of the line.

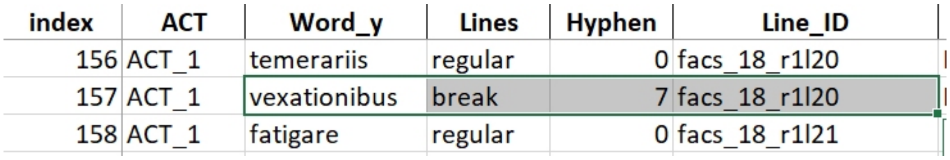
13 Col. M or O: HYPHEN

Numerical value for the cases when the graphical line is ending in the middle of the word. The value 0 (zero) indicates that the word is not divided between two distinct graphical lines. The value > 0 indicates the character position of the word where the line-break occurs.



1-19 permaneant. Decernimus ergo ut nulli hominum liceat eandem ecclesiam temere pertur¬
1-20 bare aut ejus possessiones auferre vel ablatas retinere, minuere vel temerariis vexatio¬
1-21 nibus fatigare, sed omnia integra conserventur eorum pro quorum sustentatione et gubernati¬

Figure 2: In the example the line ends at the 7th character of the word « vexationibus », in consequence hyphen value will be 7.

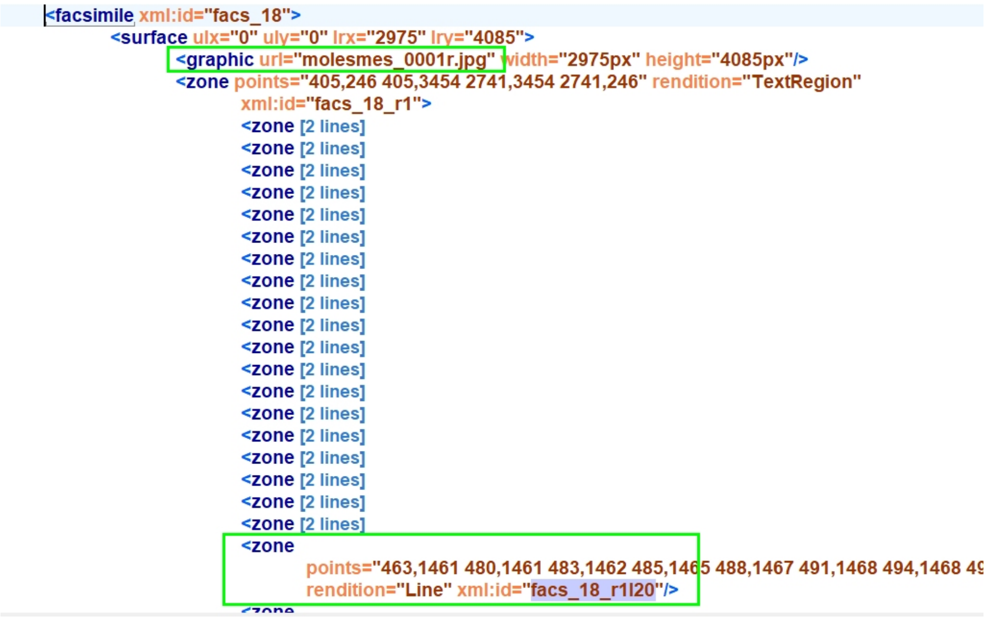


index	ACT	Word_y	Lines	Hyphen	Line_ID
156	ACT_1	temerariis	regular	0	facs_18_r1l20
157	ACT_1	vexationibus	break	7	facs_18_r1l20
158	ACT_1	fatigare	regular	0	facs_18_r1l21

Figure 3: The Line_ID of the words « temerariis » and « vexationibus » are both « facs_18_r1l20 ».

14 Col. N or P: Line_ID

The Line_ID column provides the identifier of the graphical line in which the word is located. This identifier is generated in Transkribus for each line and is unique within a document (but can repeated across several documents in one single collection). All words inside a same line share the same code. The exception is the word that carries an hyphen value >0 as the second part belongs to the next line, so hyphen value must be used in order to restore graphical line order. (see table 3).



```
<facsimile xml:id="facs_18">
        <surface ulx="0" uly="0" lrx="2975" lry="4085">
                <graphic url="molesmes_0001r.jpg" width="2975px" height="4085px"/>
                <zone points="405,246 405,3454 2741,3454 2741,246" rendition="TextRegion"
                        xml:id="facs_18_r1">
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone [2 lines]
                        <zone
                                points="463,1461 480,1461 483,1462 485,1465 488,1467 491,1468 494,1468 49
                                rendition="Line" xml:id="facs_18_r1l20"/>
                        <zone
```

Figure 4: The Line_ID identifier « facs_18_r1l20 » corresponds to an image and coordinates.

This identifier makes it possible to retrieve the name of the image and the coordinates of the line. Figure 4 shows how the image name can be retrieved in the TEI file

```
..\Database\{CodeName}\TEI\{CodeName}_tei.xml
```

here with {CodeName} being "Molesme_1_Dijon_ADCO_Cart_142_7H6"

The Line_ID is the @xml:id attribute of a <zone/> element with @rendition="Line". The ancestor::surface element contains a <graphic/> element providing the image name in the @url attribute. Coordinates on the image are provided in @points.

15 Col. O or Q: SUPPLIED

Binary value, indicates if the text should be considered (« NON ») or if it should be discarded (« OUI ») to train a model for Handwritten Text Recognition. Supplied=OUI is introduced in two situations : (i) The text found in the edition is not actually found in the manuscript or (ii) the text is present in the manuscript but it can hinder recognition since it is not well-arranged to the graphic line.

16 Optional: Col. R: RUBRIC

The RUBRIC column is not present in all files. They are only present in files with PERS_y and LOC_y annotations, therefore in column R. It consists in a binary value (OUI/NON). It Indicates if the text belongs to a rubric (« OUI »). Rubrics are a title or abstract of the text and normally placed in the first line of the charter. Not all cartularies have rubrics and not all editions transcribe them (resp. they were not available to us in a processable form). To isolate rubrics from the charter texts we have introduce a marker (« | ») at the start and at the end of the rubric and use the « OUI » value in the present column.

17 Optional: Col. P: Error_clean

Binary value. This column can occasionally be used. Here are indicates the cases in which an editorial comment that occurs in the middle of a charter was not filtered (in any case these comments are scarce and most of them were filtered). The value « NON » indicates the absence of problem, « OUI » indicates that the word must be ignored as it do not belongs to the charter text. These cases were also tagged with a « SUPPLIED » value.

Exception: in « Molesme1 », column R; in « Sommereux », column S.

18 Col. Q: Act_Bibliography

In the subset of « Île-de-France » cartularies (Chartres_1, Chartres_2, Notre_Dame_Roche, Pontoise, Port_Royal_1, Port_Royal_2), the column « Act_Bibliography » will indicate the bibliographical reference of the current act.

19 Col. R or S: Original_Act_ID

The column « Original_Act_ID » indicates the internal ID provided by the editors of the digital version. This information is displayed:

- in col. R, on the first word of each charter for the subset of « Île-de-France » cartularies (Chartres_1, Chartres_2, Notre_Dame_Roche, Pontoise, Port_Royal_1, Port_Royal_2),
- mainly in col. S, on all words of the charters for the other cartularies : Fervaques, Molesme_1, Molesme_2, Saint_Denis (col. R for « Molesme_1 »).

20 Col. S: Entity_ID

In « Nesle » cartulary, the original edition provides an identification of Named Entities through assigned identifiers. For this cartulary only, the column Entity_ID records this identifier.