

DOCUMENTARY ARCHAEOLOGY

DALME is part of the larger **Initiative for Documentary Archaeology (IDA)**, whose goal is to lay the theoretical and methodological foundations for a novel approach to the study of past material culture. We call this approach a *Documentary Archaeology* because it enables researchers to **interrogate vast datasets of textual things alongside tangible material culture** by providing a framework that translates between the **modern domain ontologies** used by scholars to characterize museum objects and archaeological artefacts and the **historical folk taxonomies** used to describe objects by people in the past.

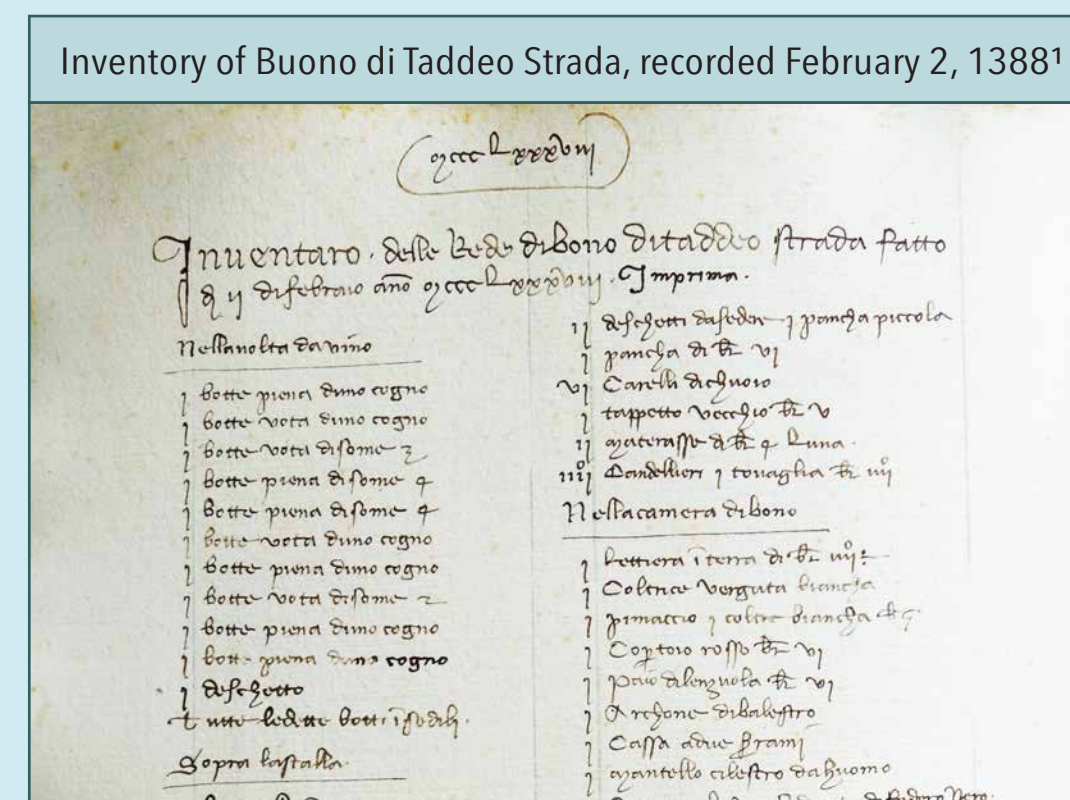
DALME DATASETS

Initially, the project focused on a corpus of late-medieval household inventories and records of debt collection from **France and Italy**. We are now in the process of augmenting our existing dataset by transcribing and incorporating additional inventories from **Toulouse, Lleida, Sicily, Florence, Bologna, Tuscany, Montpellier, Valencia** and several other cities and regions.



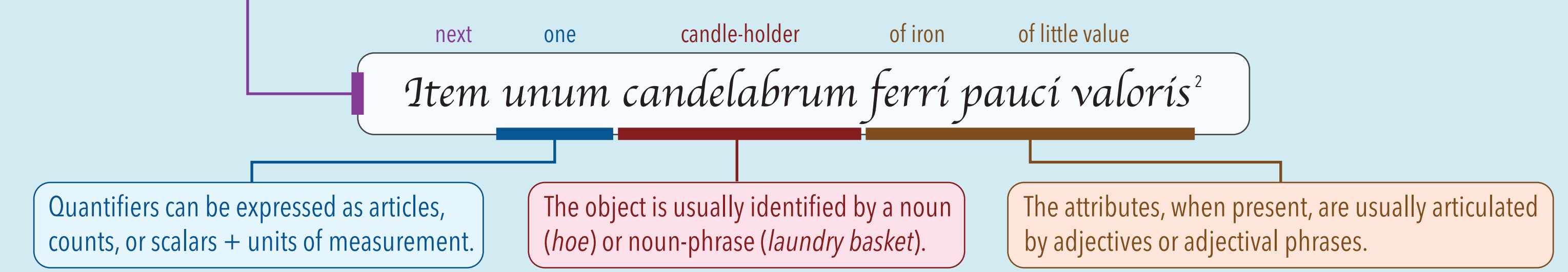
INVENTORIES

Inventories are a type of record found across a period of centuries in Europe, making it a **stable platform for serial analyses** of data across time and space. They usually represent a **large proportion of a household's contents** and derive from a **spectrum of social ranks**, capturing a far broader range of material culture than sources oriented toward high-status objects. Inventories are particularly valuable because the **systematic biases that generate silences in the textual record are often quite different from those that affect the survival of tangible things**. Comparing evidence from each of the two domains makes the systematic biases inherent in each readily apparent and subject to statistical analysis. They also offer a crucial practical advantage: as legal documents, they tend to be **highly formalized and exhibit only a small range of variation** in terms of how the data is presented.



OBJECT DESCRIPTIONS

Object descriptions within inventories can usually be separated into discrete records consisting of strings of words with well-defined boundaries (eg **order-determining phrases, such as first or next**). A record typically consists of a quantifier, the name of an object, and a list of its attributes:



OUR WORKFLOW

The project relies on a sophisticated digital architecture that helps us manage our data and documentation. Our basic workflow can be divided into three steps: **digitization, lexicalization, and semantic parsing**.

DIGITIZATION

Digitization refers to the process of generating machine-readable versions of the inventories by using **Optical Character Recognition software** (in the case of published editions) or manually transcribing them (in the case of archival materials). In both cases we use **TEI mark-up** to tag key elements of the text to facilitate importing the contents into the database as well as preserving the format and editorial marks of the document.

A

B

C

A `<list type="inventory">`

B `<item> unum coffrum ferratum cum duobus auricularibus</item>`

C `<del rend="overstrike"> frunis flunis</item></lb>`

A `<item> unum giponum modici valoris</item></lb>`

B `<item> unum manutergium</item></lb>`

C `</list>`

LEXICALIZATION

Lexicalization consists of creating the auxiliary data necessary to translate the contents of the documents. This involves **tokenizing** the contents of the documents, **lemmatizing** the words, and associating them with entries in the corresponding **lexicon** module.

CORPUS	WORD LIST	auricularibus, noun, oreiller (Du Cange)
auricularibus	auricularibus	auricularibus, noun, oreiller (Du Cange)
coffrum	coffrum	coffrum, noun, coffret (Razo)
cum	cum	cum, prep., with (Lewis & Short)
duobus	duobus	duo, adj., pl., two (Lewis & Short)
ferratum	ferratum	ferratum, noun, ferre (Razo)
flunis	flunis	fluna, noun, taie d'oreiller (Razo)
frunis	frunis	frunis, noun, japon/jupe (Razo)
giponum	giponum	giponius, noun, japon/jupe (Razo)
item	item	item, adv., besides, also (Lewis & Short)
manutergium	manutergium	manutergium, noun, torchon (Razo)
modici	modici	modicus, adj., moderate (Lewis & Short)
sive	sive	sive, conj., or if (Lewis & Short)
unum	unum	unus, prop., one (Lewis & Short)
valoris	valoris	valor, noun, value (Du Cange)

SEMANTIC PARSING

Semantic parsing consists of **reading and classifying** the contents of the inventories to generate **database records** of the objects described therein.

From the three sample entries:

A `<item> unum coffrum ferratum cum duobus auricularibus sive frunis</item>`

B `<item> unum giponum modici valoris</item>`

C `<item> unum manutergium</item>`

Five object records are generated in the database:

A `<1> COFFER style :: iron-bound`

B `<1> PILLOW / PILLOW CASE`

C `<1> JUPE condition :: of little value`

D `<1> HAND TOWEL`

CHALLENGES

A crucial challenge faced by DALME has been the development of a system that **preserves the original folk taxonomy** in the documents while also enabling us to **describe and re-classify objects according to modern research-orientated categories**. The difficulties stem from huge variability of the language used to describe material culture, both across time and space but also within the same cultural context.

For example, medieval Provençal used at least fourteen different Latinized words to refer to the object we would call *basket* in English. Furthermore, depending on context, these terms can be categorized in very different ways. While all of them refer to baskets, some can be categorized under *agricultural equipment* while others are *household equipment*. This requires the use of **flexible and non-exclusive systems of classification**, whereby each object or attribute must be classified based on its original term, rather than assimilated under an undifferentiated word (eg by collapsing the range of terms mentioned above into the English word *basket*). Conversely, these terms must somehow be classified in ways that will allow users a convenient way to query the database at multiple levels (eg pulling all the objects described by the words above when queried for *baskets*, or *containers*).

HOW WE USE LINKED OPEN DATA

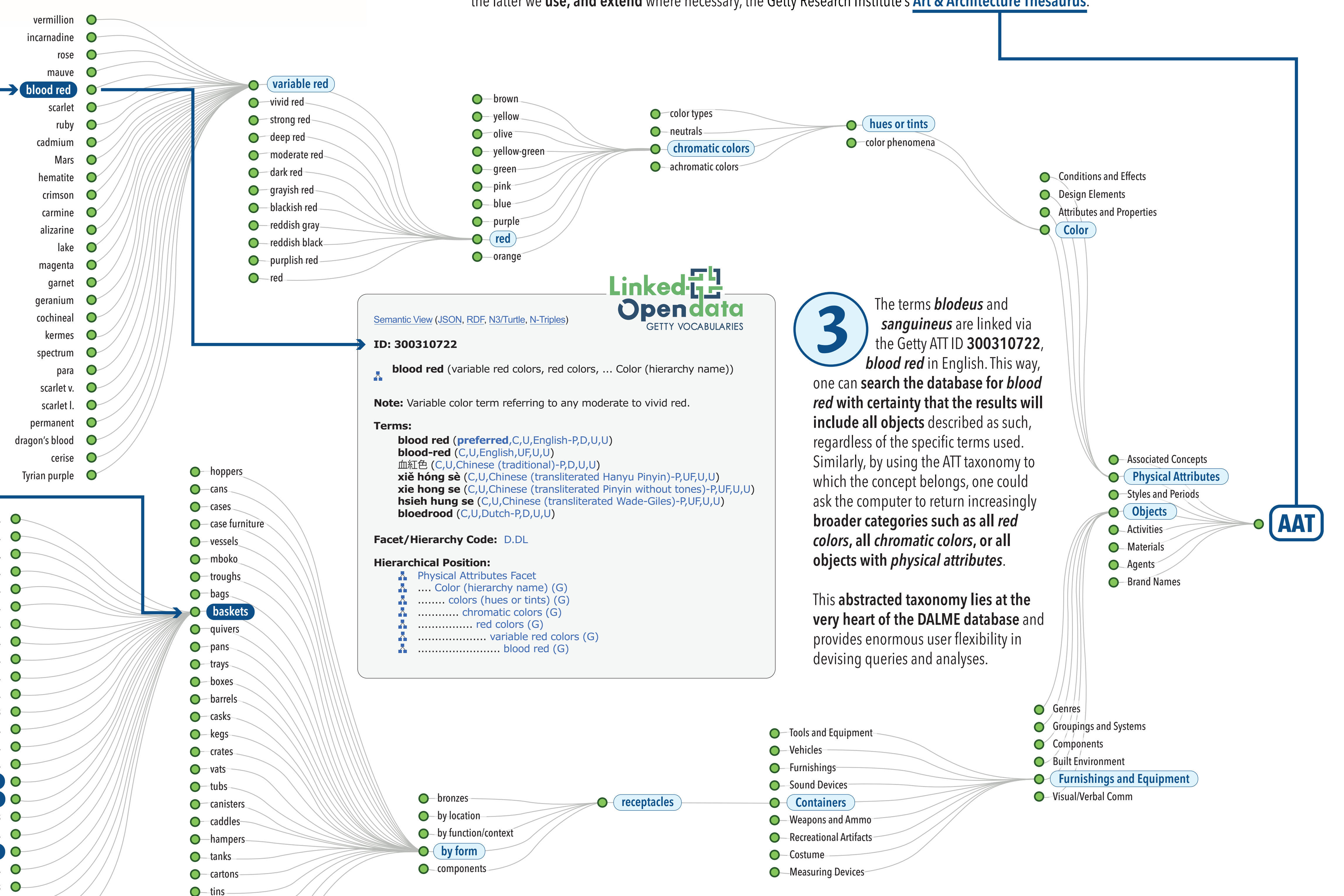
Our system uses a **faceted approach** that can assign different **scopes (temporal, spatial, formal)** to terms, in order to account for regional and temporal variations in spelling and meaning while **retaining semantic links** to the objects they describe. The latter is accomplished by defining said object as an abstract **concept** (reified only by a numeric ID in the database) to which as many terms as pertinent can be attached. As we parse the documents to detect objects and their attributes, we **semantically link** these to concepts in an **abstracted taxonomy of material culture**. For the latter we **use, and extend** where necessary, the Getty Research Institute's **Art & Architecture Thesaurus**.

2 Very different words can be used to describe attributes of an object that are in fact identical. The **two terms below are both used to describe a deep, blood red color**.

sanguineus
blodeus

4 Likewise with objects: medieval Provençal used at least **fourteen different Latinized words to refer to the object we would call basket in English**.

banasta/banastomus
bassellus
benha
belsa
canestella/canestellus
canistrum
cofinus
colla
garbella/garbellinus
gavanhola
gorbinus
panal
sportinus
systra



FUTURE DIRECTIONS: We believe that a number of other data points in our database can be managed in similar fashion to the objects and their attributes. We have recently begun to implement a similar solution for **geographic and personal names**.

SOME RESULTS

RECYCLING: Textual sources, however imperfect, provide a useful complement to archaeological evidence. In the medieval archaeological record, for example, evidence for the reuse of durable objects and materials is found everywhere. However, only some recycling processes leave behind traces seen directly in the archaeological record. Historical sources are often equally fragmentary with respect to the information they offer about recycling, but the types of information provided complement those found in the archaeological record. Explicit references to recycling objects are rare in our sources, but the frequent use of certain terms and expressions can be used to track implicit attestations of it. This graphic shows the usage of some of those terms for certain categories of material culture from inventories compiled in the city of Marseille between 1324 and 1445.

MARSEILLE 1324-1445

6,862 Objects in DALME database from Marseille

567 Objects with "recycling terms" associated

295 Objects with "recycling terms" that fall into one of the selected categories

Percentage of objects in each category with "recycling terms"

- 1 Copper and copper alloys
- 2 Iron
- 3 Wooden tableware and implements
- 4 Furniture (wood)
- 5 Ceramics (tableware, containers, tiles)
- 6 Linen (sheets, towels, tablecloths)
- 7 Covers (multiple fabrics)
- 8 Tunic (wool)

CLOTHING: Documentary sources, inasmuch as they often record organic and inorganic materials in equal measure, can provide information that would otherwise be inaccessible to archaeology. An example is provided by the distribution of fabric colors and design details in clothing. This graphic shows a preliminary analysis of items of clothing from the Italian city of Lucca between 1333 and 1342. A number of interesting patterns emerge from the dataset. If an object in the sample is gendered, it is twice more likely to be female than male. Of the materials described in the sample, all but one appear to be gender-biased towards women. Just as with materials, other descriptive terms seem strongly linked to female objects, with fringed and dyed items being almost exclusively worn by women. Lastly, looking at the distribution of colors, red appears to be the most popular color for women, while red and white seem equally popular with men. Preference for blue and green is similar for both genders.

LUCCA 1333-1342

7,489 Objects in DALME database from the city of Lucca

1,108 Objects categorized as dress or clothing

780 Objects that are gendered, i.e. stated to be either male or female

Percentage of female garments by color

- 11% blue
- 2% red
- 11% green
- 34% purple
- 18% white
- 16% yellow

Percentage of male garments by color

- 41% blue
- 26% red
- 13% green
- 11% purple
- 7% white
- 2% yellow

Distribution of gender for garments with specific materials

- silver/gold cotton
- wool
- linen
- albagia
- fur
- wool
- silver/gold = aurum, argentum
- cotton = carditatum, fustaceum
- linen = lino, linoam
- silk = sericeus, sericeus
- albagia = albagia
- fur = pelis
- wool = lanolin, camelinus, misticum, scarticus, serpis, paninus

Distribution of gender for garments described using certain categories of terms

- w/ buttons
- checked
- dyed
- w/ fringe
- lined
- parti-colored
- striped
- steeple
- w/ buttons = cum, mappalis, doctus
- checked = scarticus
- dyed = sericeus
- w/ fringe = fringed, profutatus
- lined = lineatus
- parti-colored = dimidiatus
- striped = virgatus
- steeple = steeple, masticus