From Documents to Structured Data: First Milestones of the *Garzoni* Project

Maud Ehrmann[a], Giovanni Colavizza[a], Orlin Topalov[a], Riccardo Cella[b], Davide Drago[d], Andrea Erboso[d], Francesca Zugno[d], Anna Bellavitis[b], Valentina Sapienza[c], Frédéric Kaplan[a]

[a] *Swiss Federal Institute of Technology in Lausanne (EPFL), Digital Humanities Laboratory*
CDH – INN 116, Station 14, CH-105 Lausanne.
{firstname.name}@epfl.ch
[b] *Historical Research Group (GRHis) Rouen University, France.*
{firstname.name}@univ-rouen.fr
[c] *Institute of Historical Research (IRHiS) Lille University, France.*
valentina.sapienza@univ-lille3.fr
[d] *Independent researcher*
davide_drago@tiscali.it, erboso@gmail.com, frazugno@gmail.com,

Abstract

Led by an interdisciplinary consortium, the *Garzoni* project undertakes the study of apprenticeship, work and society in early modern Venice by focusing on a specific archival source, namely the *Accordi dei Garzoni* from the Venetian State Archives. The project revolves around two main phases with, in the first instance, the design and the development of tools to extract and render information contained in the documents (according to Semantic Web standards) and, as a second step, the examination of such information. This paper outlines the main progress and achievements during the first year of the project.

Keywords: apprenticeship, early modern period, Venice, cultural heritage text annotation and processing, historical database, linked data

1. Introduction

From the Italian *garzon*, *Garzoni* is the name of these *shop boys* or *apprentices* who, during medieval and pre-modern times in Italy, were pursuing an apprenticeship with a master to learn a craftsmanship. As all over Europe, this form of training was organized and monitored by guilds. Together with cities and states, these associations of craftsmen or merchants provided the main institutional framework for the organization of labour market and, thereby, greatly contributed to shaping the social and economic landscapes of pre-industrial times. In this context, the study of apprenticeship is acknowledged as of major importance to better understand the driving forces of pre-modern economy, the transmission of skills, educational patterns and, in the case of Venice, artistic education (Gadd and Wallis 2006; Ogilvie 2011).

Who were these young boys and, sometimes, girls? Which profession were they learning exactly and under which conditions? How long were they trained and with which prospect? Possible answers to these questions — and likely many others — can certainly be found in the archival source of the *Accordi dei Garzoni* kept in the State Archives of Venice. This document series contains about 55.000 apprenticeship contracts declared by guilds of various professions over an almost continuous

two-century period (1575-1772)[1]. The working arrangements contained therein represent an important source of information to understand, amongst others, who was working where, for how long and how much, how workshops were organized and distributed over the city, which professions were taught and where apprentices were coming from. The objective of the *Garzoni* project is to devise computational tools and methods so as to fully exploit the potential of this source and to support the advancement of knowledge about apprenticeship in the early modern Venice. To this end, the project revolves around two main phases with, in the first instance, the development of tools to extract information contained in the documents and, as a second step, the examination of such information. Even though distinguished, these phases are closely interdependent: while the design of computational tools needs to be informed of historical needs and constraints and to cope with historical material, enquiries on the collected data need to confront with both the limits and opportunities of computational methods and/or computationally acquired data.

In this context, the first milestones of the project address the tasks of collecting, processing, storing and rendering information contained in the archival material. However, if automatic document processing techniques have achieved a certain maturity for documents of present time, the transformation of hand-written documents into well-represented, structured and connected data which can satisfactorily be used for historical study purposes is not straightforward and requires several processing steps. Transitioning from documents to structured data is one of the key challenges facing the *Garzoni* project.

In this mid-term project statement, we report on the main advances during the first year of the project, focusing particularly on the issues of how to represent, extract, enhance and exploit information contained in the archival material while meeting – and maybe triggering – the needs of historical studies. The remainder is organized as follows: after introducing the document series of the *Accordi dei Garzoni* (section 2), we present the rationale and architecture of computational processes (section 3). We then follow the course of the different computational steps foreseen to support historical research questions (section 4, 5, 6 and 7) before having a first outlook on the envisioned explorations of the acquired data (section 8). We finally give institutional details (section 9) and conclude and consider future work (section 10).

2. The *Accordi dei Garzoni*

The *Accordi dei Garzoni* (ADG) is a document series from the State Archives of Venice which originates from the activity of the *Giustizia Vecchia* magistracy. By the enforcement of laws from the 13th and 14th centuries, this judicial authority was in charge of registering apprenticeship contracts with the aim, among others, of protecting young people while they were trained and/or providing domestic services (Bellavitis, 2006). If all masters and guilds did not always comply with the legislation – although regularly reiterated –, the result of this regulation is that information for much of apprenticeship arrangements got centralized, today reflected in an exceptionally dense and complete archival series.

The *Accordi dei Garzoni* comprises 32 registers which are, for the most part, in a very well-preserved state. The first register (n°151) starts on June, 9th 1575 and the last (n°182) ends on May, 20th 1772. Despite some gaps at the beginning and at the end of the 17th c., the coverage of the bound period is pretty good and contract records add up to ca. 55,000. Registers contains 3 to 6 records per page, each amounting to a small paragraph of 6 to 10 lines preceded by the date. An additional note would sometimes appear in the margin, indicating a correction or modification made to the contract a posteriori. Apprenticeship enrolments were registered by several officers, resulting in very different handwritings. Deciphering such writings is nowadays restricted to experimented

[1] Venice State Archives (ASVe), *Giustizia Vecchia, Accordi dei garzoni*, bb. 112-126, registers 151-182 (1575-1772).

paleographers, specialists of Venetian dialect and familiar with the numerous abbreviations used by the scribes (cf. Figure 2).

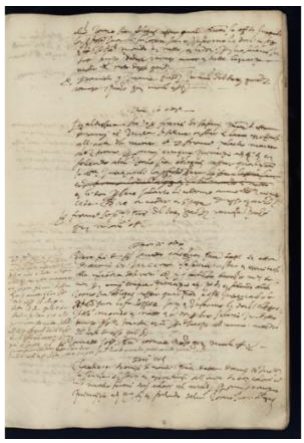

Fig. 1. Accordi dei Garzoni, box 115, register 158, page 18r.

The vast majority of contracts were recorded following the same pattern, that is to say documenting the same elements of information. Let us consider the example of *Baldisera de Zuanantonio*, which contract is shown in Figure 3 below.

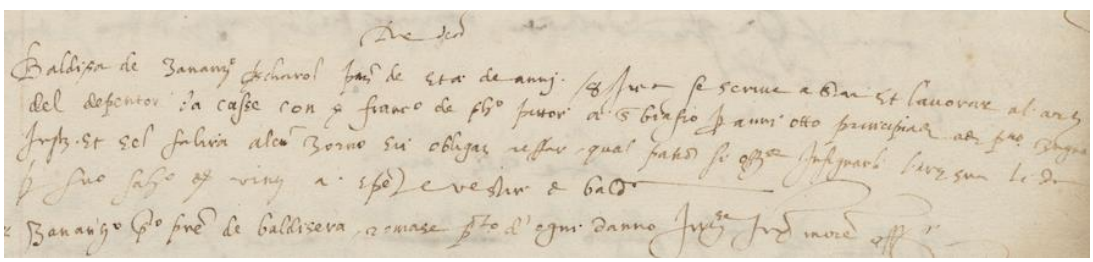

Fig. 2. Contract of Baldisera de Zuanantonio box 112, register 151, page 18r.

Within these few lines, the essentials of the apprenticeship agreement are stated, with:

- the identity of the involved persons, with the Apprentice, the Master, and potentially a Guarantor; here *Baldisera de Zuanantonio* commits to work with the master *Francesco de Philippo*, whose agreement is guaranteed by the father of the boy, *Zuanantonio*. Information such as the age (here 8), the residence or the geographical origins of the apprentice are often mentioned.
- the profession to be taught and the workshop where the learning takes place; here *Baldisera* will learn how to paint chests, in a workshop located in the parish of *San Biasio*.
- the diverse terms of the contracts, with the salary (here 20 ducats), the mention of possible advantages in kind, the contract duration (here 8 years), and at times other details.

The regularity of the provided information is a major advantage: it allows a systematic study of the contracts and greatly eases the definition of a data model. But how to exploit such material? How to collect this large amount of information and make it accessible to everyone? The next section details the processes we devised to enable the transition from documents to structured data.

3. From documents to structured data

Our starting point is tens of thousands of Garzoni contracts and our finishing line is an information system to make sense of all the information they contain. Starting from digitized documents, processes are incrementally organized into the following steps, where information is gradually built up:

1. data modelling – or how to formalize what we are interested in;
2. data acquisition – or how to extract information, via transcription and annotation, according to the data model;
3. data processing – or how to disambiguate entities, perform inference and interlink with other data-sets;
4. data exploitation – or how to enable search and visualization.

The first step, the definition of a data model, corresponds to the identification of entities, concepts and relations which are of interest (step 1, section 4). Once equipped with a model, it is possible to proceed with the transcription and annotation of the documents, in order to acquire data and populate a database (in our case a RDF graph) based on the model. To this end, we develop a web-based annotation and transcription interface, named *DHCanvas* (step 2, section 5). At the end of this annotation phase, a certain amount of data is collected, which is however not yet exploitable because of linguistic variation (especially person and profession names) and of entities pertaining to different levels of genericity/specificity (especially location names). Disambiguation and interlinking strategies are thus necessary (step 3, section 6). Finally, it is important to enable non computer-scientists to visualize and search through the acquired data; this is the role of another web-based interface, named *DHExplorer* (step 4, section 7).

The expected outcomes of these developments are new ways of addressing historical questions with:

- access to more data (*n* = all contracts of the *Accordi dei Garzoni*);
- possibility to ask complex questions (e.g. the average salary of apprentices originating from city *c* working in profession *p* from date *x* to date *y*);
- possibility to quickly test hypothesis;
- possibility to imagine new questions, as the data itself can generate the question;
- simply and most of all, possibility to access data, because openly and online delivered.

The four steps previously defined and presented hereafter — data modelling, acquisition, processing and exploitation — establish a workflow which is not specific to the Garzoni project. Other cultural heritage 'data mining' endeavours could be similarly framed, each step being tuned to particular needs and constraints.

4. How to represent knowledge and store information

The definition of a data model corresponds to the identification of what is of interest in the 'world' — in our case *Garzoni* contracts — we are looking at. Regardless of the format, the role of a model is to represent the structure and the main elements of a domain in order to explain and/or dynamically reproduce it. As a result, a model should enable to answer questions or solve issues regarding the modeled world. The modelling step is thus a crucial one: it establishes the work-space and prepares what can be done.

To represent a portion of reality we need to identify its concepts, its entities, their properties and their interactions. In the context of *Garzoni*, we are quite fortunate for the content of the contracts — recall it is administrative documents — is extremely regular, with always the same pieces of

information. Before starting off with the definition of the vocabulary, we first had to address two issues: the choice of a formalism and the characterization of the scope of the model.

4.1. Representation language

Regarding knowledge representation formalism, we chose to express our model through an ontology, using the OWL (Web Ontology Language) language. This choice is motivated, among others, by the wish to (a) exploit inference capacities, (b) re-use domain knowledge, (c) benefit from the technical eco-system of the Semantic Web and (d) integrate and share our data according to Linked Data principles (Auer, 2011; Heath, 2011).

4.2. Scope of the model

As for characterizing the scope of our model, a pre-study of the documents coupled with the consideration of historical research questions at stake in the *Garzoni* project allowed us to formulate competency questions, that is to say questions that we wish the knowledge base could answer. In the context of *Garzoni*, research questions revolve around economic aspects (e.g. how did apprentices' salaries differed according to professions, guilds, time, etc.), knowledge transmission (e.g. were apprentices just a source of labor or were they trained), migration flows (e.g. where did apprentice come from), family and gender issues (e.g. at what age boys and girls were sent away to become apprentice), and others. Each question can be looked at from a strict Venetian view point, or put into the larger European perspective[2]. Computer scientists together with historians worked on the formulation of these competency questions while taking into account their implications with respect to the data model.

Besides, we also benefited from a sub-set of 9000 thousand annotated contracts (out of the pre-study phase) from which we could derive evidence regarding major features to take into account or not. An example is the use of the property *nickname* for a Person which, after believing it was of importance, turned out to be extremely rarely used. We encountered a similar situation with the encoding of the information of *pledges*, which were sometimes given as security for the fulfillment of the contract. This case occurred only 20 times out of 9000, but was deemed worth keeping. Here arises the issue of finding the right trade-off between representativity and effectiveness; all details cannot be kept, or we are at risk of having an unmanageable model (on which several tools have to be built), but all details cannot be dumped, or we might end up with an unrepresentative and meaningless data-set. The capacity to find the correct trade-off naturally depends on both computer scientists and historians; in our case we "negotiated" on a case by case basis, giving priority to detail inclusion when encountering borderline cases.

At this point it is worth having in mind that an ontology is inherently relative: it corresponds to one of many possible views of a domain (which differ according to time, cultural background, targeted application, etc.) and it evolves through time. Ours is naturally shaped according to our needs, and is iteratively revisited.

4.3. Garzoni data model

Even though the model might evolve further in the future, it can be seen as fairly stable as of today. Our intention here is not to detail the whole ontology but rather to emphasize some points of interest and to present its main building blocks.

---

[2] Regarding the European perspective, the *Garzoni* project would wish, on the long term, to cross data-sets and share studies with projects of similar vein.

4.3.1. Generic modeling issues

Before describing the core content of *Garzoni* data, we first focus our attention on key modeling questions, namely (1) how to represent information which belongs to different levels, (2) how to represent information which evolves through time and is dependent on a specific view point, and (3) how to trace its provenance.

Mentions and Entities

When dealing with textual data, an obvious aspect to model is the difference between what appears in the text (the word) and what it is being referred to (the 'thing'). In the case of lexical information, this amounts to distinguish between an occurrence of a word in a text and its meaning as stated in a dictionary. When dealing with entities, this distinction corresponds to the *mention(s)* of an entity in texts on the one hand and, on the other, the *entity of the world* they refer to. In our model, we therefore introduce the distinction between *entity mentions*, at the document level, and *entities*, at the world or entity level. The former is used to represent the various possible occurrences or mentions of e.g. *Francesco de Philippo* in different contracts, and the latter to represent the unique person of Francesco de Philippo who lived at a certain time. Similarly, the Venetian quarter *Dorsoduro* might be mentioned thousand times in the registers, but it always refers to the same 'Sestiere' of Venice (which might itself have endured changes through time, but this is another point). The link between the two classes is reified, in order to represent meta-information about the relationship itself, especially which agent led to its creation, a human or a machine (as of now simply represented by a Literal[3], i.e. a string). Figure 3 displays the organization of these classes. In the future, this will allow an informed exploitation of the data.

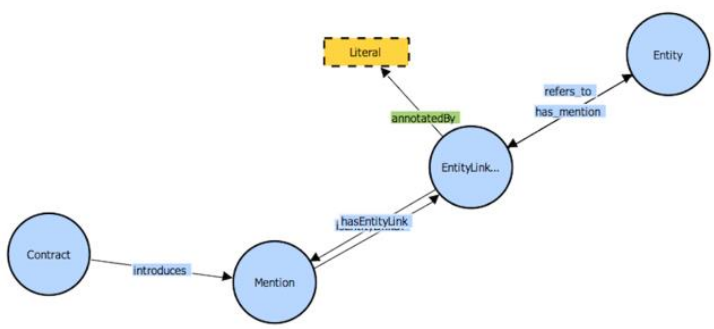

Fig. 3. Entity-Mention representation

Statements

In our model, entities (not mentions) are objects which collect information about different states of existence. No matter how this 'snapshots' will be collected, they all originate from a given source and assert some properties valid at a certain time. Without considering here all the options possible to encode this information, let us just mention our choice, that is to say the usage of *Statements* similarly as in the RDF representation of Wikidata (Vrandecic, 2014). Each entity has a set of property-value pairs, whose values are not the value itself, but a *Statement* giving access to the value, its validity time, and its source or reference. If the entity *Francesco de Philippo* appears in two contracts, first as a Master, second as a Guarantor, it will have two role statements, asserting that

[3] https://www.w3.org/TR/rdf11-concepts/#section-Graph-Literal

this person was a Master at a certain date according to a certain source (a contract), and a Guarantor at another date according to another source.

Provenance

Garzoni data is acquired through a web-interface (cf. section 5) whose internal model relies entirely on the Web Annotation Data Model[4] (OA). Each and every instance correspond to an annotation which relates a target (what is annotated) and a body (the content of the annotation) and is further qualified by the motivation of the annotation and the agent who created it. Concretely speaking, this means that each mention or entity is the body of an annotation, whose target is, in the case of a mention, an image segment and, in the case of an entity, a mention. The internal annotation scheme of the web-interface — not further detailed here — therefore constitutes a meta-information level in which the garzoni data is "wrapped". Based on this input, the *Garzoni* RDF data-set will naturally keep this provenance information.

4.3.2. Garzoni content core elements

The Garzoni data model covers different areas of information, which can be briefly described as follows (cf. Figure 4): A Contract has an Archival Signature, introduces Entity Mentions and reports about Events and Conditions (contract terms).

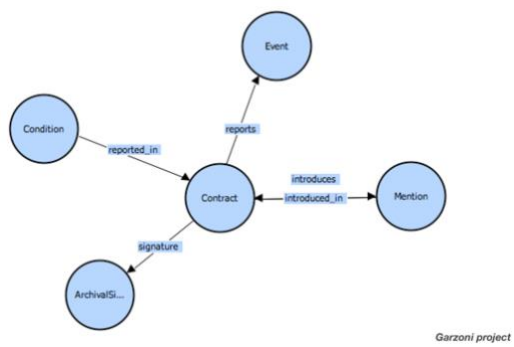

Fig. 4. Basics elements of Garzoni model: A Contract has a Signature, introduces some Entity Mentions and reports about Events and Conditions.

Information about the sources (archival dimension)

A garzoni contract (a document) is materialized on two mediums: the real paper registers on the one hand, theirs scans on the other. For both mediums, the ontology represents appropriate descriptive and technical metadata information. Figure 5 below illustrate this part of the ontology; further aspects about the digitization process are described using the CRM-dig ontology (not represented in the figure) (Doerr, 2011).

Information about various entities or concepts:

Garzoni contracts feature different actors, with different roles (Apprentice, Master and Guarantor) and characteristics. Contracts are quite consistent and the mention of a person follows more or less the same pattern, with the indication of his/her name, age, profession, residence, geographical origins and potentially his/her charge (a kind of public responsibility, different from

[4] https://www.w3.org/TR/annotation-model/

the profession). These attributes are however not evenly distributed: the age is mentioned for the Apprentices exclusively, while charges are often indicated for Guarantors. Besides Person, contracts mention Places (sometimes with the indication of the Sestiere — a

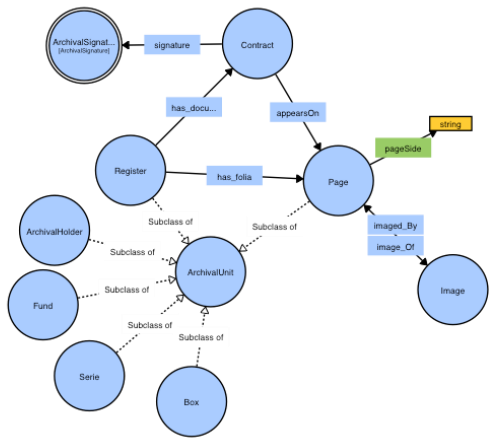

Fig 5. Part of the model related to archival aspects.

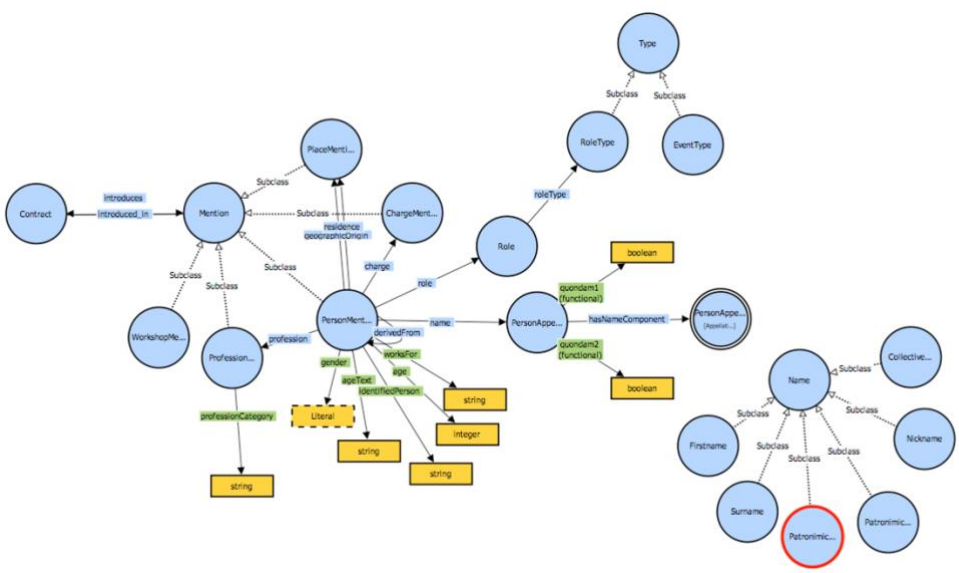

Fig. 6. Entity mentions of type Person, Place, Charge, Workshop and Profession, along with their properties.

Venetian district — and of the Parish), Workshops (with their *insegna*) and Professions. The ontology has classes and properties to represent all information about these entity mentions (cf. Figure 6).

Information about contracts' terms (condition dimension):

Each contract gives information about financial conditions and what we called 'hosting' conditions. The first ones relate to the salary and/or the pledge which were due by one of

the parties. Salaries can be quite complex, with different amounts paid by stages. Hosting conditions relate to various aspects of life which were sometimes supported during the apprenticeship: the general expenses of the apprentice, his/her clothing, personal care and accommodation. Financial and hosting conditions are defined classes in the ontology, made up of some condition components (salary, accommodation, pledge, clothing, etc.).

Information about events

Garzoni contracts give evidence of the occurrence of several events. The apprenticeship itself, its enrolment at the *Giustizia Vecchia*, the potential breach of the contract as well as the apprentice's flee (and, sometimes, his/her return). To represent this information, we make use of classes and properties of the Simple Event Model (SEM) ontology (van Hage, 2011).

Information about relations

Relation can occur between contracts (a contract cites another one) or between persons (mainly family and business relations). They are stated between entities.

This first version of the *Garzoni* data model covers the identified needs in terms of content representativity, traceability and manageability, and is used by the web interface presented in the next section. With the exception of CRM-dig, SEM and OA, the ontology is however mostly expressed via 'in-house' classes and properties; it is our intention to further link our vocabulary later in the project (Janowicz, 2014). Finally, it will be published online in a human-readable version[5] and with appropriate content negotiation.

5. From images to structured content

We wish to extract information from *Garzoni* contracts, that is to say to transcribe and annotate them according to our data model. Given the variety of handwritings present in the ADG and their time-span which necessarily affect language and abbreviation, using an automatic handwritten text recognition software is not an option. Indeed, although handwritten text recognition for historical texts became a very active field of research in the past few years[6], performances are not yet satisfying and quite some archival material will probably remain out of the reach of algorithms for some more years. (Additionally, automatic transcription will produce a text, which still would have to be annotated afterwards.)

Instead, we develop an innovative web-based application for visualization, transcription and annotation of historical documents, named *DHCanvas*. Its overall goal is to allow users to collaboratively transcribe and semantically annotate documents, thereby producing information which can be stored, processed, retrieved and exchanged. Some of the key features of DHCanvas are the preservation of the link between the annotation (already an interpretation) and the source (an image segment), the capacity to handle concurrent annotations and its full compatibility with representation standards for digital resources (IIIF[7], Shared Canvas Model[8]) and annotations (Open Annotation Data Model).

Even though application scenarios involving image annotation can be diverse, we currently focus on annotation of images of textual documents. In this scenario, the user reads a page, selects an image segment (usually corresponding to a segment of text) and is being offered the possibility to

[5] Using the LODE environment, http://www.essepuntato.it/lode

[6] See for example http://transcriptorium.eu/~htrcontest/

[7] International Image Interoperability Framework, http://iiif.io/

[8] http://iiif.io/model/shared-canvas/1.0/

transcribe and annotate the textual chunk imaged by that segment. The annotation of a segment involves several sub-annotations which encode various types of information. In a document annotation context, information can be regarded as belonging to three different levels:

- resource level — what relates to a digital resource/canvas, e.g. an image segment representing the name of a person;
- document level — what relates to textual content, e.g. an annotation representing the mention of a person;
- entity level — what relates to concepts and entities of the world, e.g. an annotation representing the entity a mention refers to.

Fig. 7. DHCanvas annotation space on a Garzoni image segment.

In concrete terms, the above-mentioned levels are reflected by the annotation workflow offered to the user in DHCanvas. At level (1), the user *selects* (or draw) a segment of the image and *transcribes* it. The *transcription annotation* consists of the literal transcription (mandatory), the language used (mandatory), the certainty of the transcription (optional) and the standard form to which corresponds the transcription (optional). An example would be an image of a Garzoni contract, with a segment featuring the hand-written name "Francesco de Philippo".

At level (2), the user *describes* what the text is about, that is to say creates a *mention annotation* gathering *local* information about the concept or the entity *mentioned* in the current segment. This information consists of a semantic category (the broad category of things to which the mention belongs to), a tag (a kind of sub-category which specifies the role of the mentioned entity within the local context) and a series of attributes. Regarding our previous segment, a *mention annotation* can be created for the segment "Francesco de Philippo", with the category *PersonMention* and the tag *Master*. If present, other attributes can be specified, e.g. his profession or residence (cf. Figure 7).

Finally, at level (3), the user *identifies* to which real-world entity the mention refers to, that is to say creates an *entity annotation* meant to identify and gather information about the real-world concept or entity, independently from the local/textual context. The information associated to the entity consists of a name, which should be unique, and possibly a relationship towards another entity.

In our case a *Person* entity is created, it corresponds to the unique person of Francesco de Philippo and has the identifier "Francesco de Philippo (pittore)".

In this way, information about an entity is progressively assembled by collecting elements attached to related mentions. The entity "Francesco de Philippo (pittore)" collects information from diverse sources with different temporal and viewpoints profiles. It is possible to encode the knowledge that according to contract *a* at time *x* Francesco was a master, that according to contract *b* at time *y* he was a guarantor, and that according to such tax declaration (coming from another document series) at time *z* he was married to this or that person.

Technically speaking, information is stored in DHCanvas database (PostgreSQL, with extensive usage of JSON schema) before being exported to RDF. Because of the inner data model of DHCanvas, which uses standard vocabularies and the JSON format, the export operation is very light.

DHCanvas annotation capacities are naturally based on visualization and image manipulation functionalities. These consist of passive visualization functionalities (e.g. one-page display, pan zoom, rotation, filmstrip, navigator) and active visualization functionalities (with impact on data, e.g. create, resize, group and delete a segment). Upcoming functionalities include collection and book mode views (i.e. several pages) as well as linking, merging and drag-selecting segments.

The DHCanvas additionally supports search functionalities: a user can look up a specific string, instantaneously retrieve all image segments where the string appears, and finally jump on the exact page where the string was mentioned. Future features of this powerful tool include the integration of complementary services via REST APIs, vocabulary definition and management by user, metadata management (for descriptive metadata), and user management, including user role and permission management.

Work on DHCanvas is on-going; when mature enough the application will be duly documented and the code shared on the web-based Git repository hosting service GitHub. Regarding the timeline within the *Garzoni* project, the annotation campaign is starting with the second year of the project. Experimented paleographers who are already familiar with the ADG series start to annotate using the DHCanvas. In order to ease the annotation task and the usage of the interface annotation guidelines are provided; constant feedback is shared between the teams. At a later stage, we will carry out an evaluation of the annotation to evaluate inter-annotator agreement on the different tasks (transcription, classification, attribute assignment and entity identification). The objective will be to estimate the difficulty of the task, to value the functionalities of the interface and, finally, to assess the quality of the material historians will use.

6. From content to enhanced information

Transcription and annotation with DHCanvas enable the transition from images of documents to structured content. This is a giant step which offers the possibility to explore information in an unprecedented way, embracing the detailed content of all documents. However, as mentioned earlier, linguistic variation, ambiguities and concept expression at different levels of specificity still hinder a full-fledged use of data. We first list the different processing steps we anticipate before describing in more details first experiments we conducted for entity record linkage.

6.1. Enhancing information

A first crucial step to meaningfully exploit *Garzoni* data is the building of a taxonomy of professions. A hierarchical classification of profession concepts is indeed necessary to abstract away from the several hundred profession names and their variants. The challenge here is that even if

historical occupation databases exists (e.g. HISCO[9]), whose classification could be projected on our data, professions which appear in Garzoni contracts are in Venetian dialect and, quite often, unknown. Our approach to cope with this is to adopt a hybrid strategy where historians devise a preliminary taxonomy based on complementary archival sources and on the empirical evidence of early annotated professions, this process being iteratively repeated until exhaustion of profession names. Quickly after the first round of this process we plan to integrate a first profession taxonomy within the DHCanvas so as to annotate appropriately already classified professions, while others will reincorporate the taxonomy building process. As shown in Figure 8, the more contracts are annotated, the more the apparition of new profession names decelerates. Although this could only be verified in the future, we hypothesise that a few iterations might be necessary to build a taxonomy covering all professions. This work will be helpful not only to query data more efficiently, but as well to deepen historical knowledge about professions and their evolution during early modern times in Venice. Ultimately, when possible we will link profession names to HISCO identifiers.

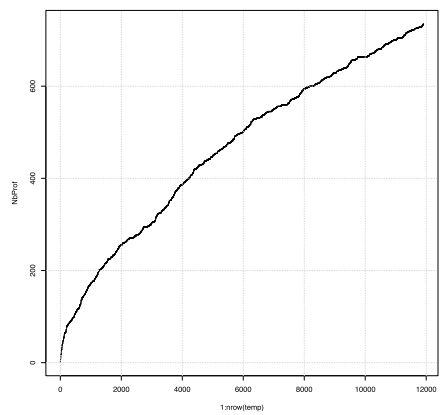

Fig. 8. Number of unique profession names (y-axis) in relation with number of annotated contracts (x-axis).

Another processing that can enhance data is inference, or the process to build new information on the base of known information. To this end, we intend to use the reasoning capabilities of the OWL language to infer new statements using our ontology, especially about person relations. Either programmatically or via rules, we are already able to locally complement information e.g. to calculate the theoretical ending date of a contract based on his starting date and duration, or to establish a master-apprentice relation between persons having these roles and involved in the same contract.

Finally, interlinking Garzoni data with other data-sets (e.g. DBpedia[10] and Geonames[11]) will be of great benefit, for both complementing and better structuring our data.

6.2. Record linkage with sparse historical data: first experiments

Major challenges when dealing with people-related data are homographic person names referring to different persons, as well as the existence of name variants referring to the same person

[9] Historical international classification of occupations, https://socialhistory.org/en/projects/hisco-history-work.
[10] http://wiki.dbpedia.org/
[11] http://www.geonames.org/

(Bloothooft, 2015). These are well-known issues in the field of Natural Language Processing for which various approaches have been devised, first via mention clustering (Mann, 2003; Artiles, 2008), more recently via linking to a knowledge base (Ji, 2011; Shen, 2015). Although closely linked and often used interchangeably, we distinguish between *record linkage*, the objective of which is to match co-referring textual mentions (based on mention contexts), and *entity disambiguation*, the objective of which is to link one or more mentions to a unique identifier in a knowledge base (based on mention and entity contexts).

The *Garzoni* database will contain a large number of person and places mentions and will therefore be in need to perform record linkage. This task is particularly difficult with historical data for several reasons: (i) its sparsity, meaning the ways of mentioning a person and the mention context are varied thus the overlap space between any two mentions might be very small; (ii) its irregularity, both orthographical and morphological; (iii) its intrinsic wide disambiguation space, meaning that several times the available information is insufficient to unambiguously link two mentions, but a probability distribution over possible mentions must be given in turn. The added challenge with the *Garzoni* database is its dynamics and diplomatic aspect. Dynamics means evolution through time, by which masters can be mentioned several times over their career (easy disambiguation task) or apprentices could become master later on (more difficult task). Diplomatics means a formulary way to mention each person in a contract according to his/her role. As an example, apprentices are usually mentioned with their age and place of origin, features which are rarely, if ever, provided for masters and guarantors.

A system for record linkage has been developed in order to be embedded as a recommendation system within the DHCanvas (Colavizza et al., 2016). The proposed approach is built on the combination of three record linkage methods: a comparison of the features of each person mention (such as name, surname, age, profession, etc.); the presence or absence of a combination of features characteristic of a specific role in a document (e.g. age and place of origin for apprentices); and the similarity of core features, such as the given name, which acts as a filter on the disambiguation space (all the pairs of mentions to be compared for disambiguation). The proposed method reaches a precision of 61% for mentions of masters and guarantors—the mentions of apprentices are intrinsically harder to disambiguate, as rarely an apprentice is mentioned in more than one contract, either as an apprentice or, in the future, as a master.

7. How to visualize and access information?

The *Garzoni* data-set will be available through a SPARQL endpoint, dereferenced URIs as well as a RDF dump. However, in order to ease the exploration of the data by non-experts, we additionally develop an interface to visualize and search through it, named *DHExplorer*. This interface is complementary to the search at segment level offered by DHCanvas; on the long run, the two interfaces are meant to be connected.

DHExplorer is a web-based interface which allow to explore the Garzoni linked data-set according to three perspectives: the *entity view*, to look closely at individuals, the *network view*, to see how people are connected and the *statistical dashboard view*, to get an overview of different aspects of the data. More specifically, the statistical dashboard offers graph representations allowing to quickly get elements of answer to various questions, e.g. the competency questions. It is divided among the following sub-views:

- Archives – to explore e.g. the distribution of contracts over years or per register;
- People – to see e.g. the top-mentioned masters, or the age distribution of apprentices according to a certain time window and/or a profession;
- Economy – to understand e.g. in which profession the salary is paid by the master or by the apprentice;
- Event – to see e.g. the average duration of contracts, the distribution of apprentice flees over the years;

- Geography – to visualize the distribution of workshop per Sestiere and Parish within Venice, and the geographical origins of apprentices.

A prototype of this interface already exists but its development is on-going; it will be put to the test when data will be available.

8. Studying Venetian apprenticeship through the *Garzoni* database: the economic history perspective

Several long-lasting historiographical questions converge around the institution of early modern apprenticeship. Economic historians are interested in the ways the institution of apprenticeship, and its direct parent guild, worked across time and space. This debate is currently polarized between who regards the guilds as being at the core of the process of transmission of (practical) knowledge as well as welfare for orphans and youth in general (cf. Epstein 1998); and who instead regards guilds as an exploitative institution, whose main focus was safeguarding members' privileges and allocating their dividends (cf. Ogilvie 2014). Naturally, a concrete appreciation of the apprenticeship in early-modern Venice, a domain for which there exist no comprehensive study (Bellavitis 2006, p. 50), would greatly contribute to this debate.

The interests of economic historians are not limited to the role of guilds in the transmission of knowledge but also span into the organization of economic activity, the dynamics of the Venetian economy and the interplay of different activities over the two hundred years considered by the *Garzoni* project. It is well known that the 16th century saw perhaps the last period of Venetian activity at a global economic level (cf. e.g. Rapp 1976), followed by a century of alleged crisis, or for some transformation and adaptation to a more marginal, regional role (cf. Sella 1997). Lastly, the 18th was a century of increased ferment in new-found or rediscovered sectors, also due to the growth of industrial activities over the mainland. It goes without saying that following the dynamics of apprentices in the dominant city of Venice might serve as a proxy to understand and map, at a finer grain than ever before possible, those more global trends.

Preliminary investigations have already highlighted how the wealth of data from the *Garzoni* project might contribute to the current historiographical debate. As an example, Colavizza (2016) considered a sample of contracts from three periods of importance for the Republic of Venice: 1) the end of the 16th c., marked by the so called plague of Saint Charles (1576), the war of Cyprus and the famine of the end of the century; 2) the '20s of the 17th c., just before the next major epidemic outbreak (1629-30); and 3) the '50s of the 17th c., as the late stages of the war of Candia took place. Two questions were addressed: (1) do the *Garzoni* data follow the trends of the economy of Venice as we know it from previous historiography? (2) Is it possible to give a first characterization of the concrete nature of Venetian apprenticeship: was it used as a source of cheap labor (leaning towards the interpretation of Ogilvie) or as a means for knowledge transfer to youths (as Epstein suggests)?

The proxy of apprenticeship contracts shows signs which could be interpreted as an increased contraction or reorganization in the economy of Venice during the three periods under analysis—confirming previous results (e.g. Rapp). Whole sectors virtually disappeared, like the printing press (as already highlighted by previous studies, see Infelise 1997). In general, as previously noted too (Lanaro 2008), several signals point to a less varied and increasingly more introvert economy over time. If we look at the proportion of registered apprentices born in Venice, as opposed to those coming from outside of the city, we see the following progression: 34% for period 1, 46% for period 2 and 56% for period 3. As a confirmatory corollary, the proportion of contracts ended due to the early departure of the apprentice (who fled) consistently decreased from 15% in period 1, to 10% in period 2 and 5.8% during period 3. The rise in the proportion of Venetian apprentices is even more surprising given the huge influx of immigrants following the great plague of the years 1629-30,

which allowed the city to rapidly recover to almost its pre-plague population (cf. Alfani 2010). Another possible sign of the contraction of the 17th c. is the reduced variety of registered professions, which can be appreciated in Fig. 9. "Essentially, the same number of apprentices were directed towards fewer professions in variety, and at the same time fewer professions were getting the lion's share of fresh recruits" (Colavizza 2016, p. 4). To be sure, several possible explanations are currently possible, given the preliminary status of our project. It might be equally likely that changes in the regulation of specific guilds, as well as into their record-keeping practices might explain the same phenomenon.

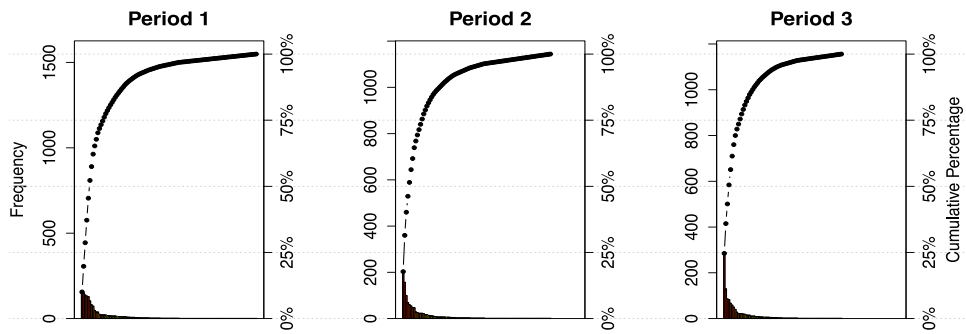

Fig. 9: Number of contracts for all professions over the three sample periods. Note the reduction of the distribution tail over periods, in terms of both its length (professional variety) and 'fatness' (professional concentration).

On the other hand, analyzed through the lens of its 'inner mechanics', venetian apprenticeship shows remarkable regularities. Apprentices can be divided in two well-defined groups: who received a salary from his/her master (roughly 80% of the registered apprentices) and who did not receive a salary or paid the master in turn (the remaining 20%). Venetians by origin had a preference for *not* being paid a salary: if a mean of 44% of the apprentices over the three periods were Venetian, 24% of these did not gain a salary while only 11% of the non Venetians gained no salary, on average. These proportions follow the increase in the number of Venetians over the three periods. Such preliminary evidence might hint at a distinction between two groups of apprentices: who gained a salary was mostly occupied as labor force, who did not gain a salary or paid the master in turn was more likely to received training over the period of apprenticeship. This distinction was also likely to be related to different crafts (Bellavitis, 2006) and can be compared to other European contexts (De Munck, Kaplan, Soly, 2007). For example, the case of London highlights that rising premiums paid to masters were linked with younger apprentices, deemed less productive (Wallis, Webb and Minns 2010, p. 400), and lower premiums were due to their masters by apprentices with previous «exposure to the occupation» (Minns, Wallis 2013, p. 347).

This working hypothesis must be further analysed during the project, yet further evidence points to it. Considering only the approximately 80% of the contracts with a salary paid by the master to the apprentice, it is possible to investigate some factors which influenced the amount of salary paid. Using standard regression analysis techniques, the salary is shown to be strongly correlated to the age of the apprentice and inveresly correlated to the length of contracts; given that very rarely we find apprentices taking two or more apprenticeships, this result can hardly be explained by previous experience into the craft. Therefore this pattern could be linked to common labor market mechanics where the added value is for example physical fitness, but not previous experience into the craft: older apprentices were paid more and serving for a shorter time because they were more fit as unskilled labor force. To be sure, other possible explanations exist and will be verified in due course,

such as the existence of previous relations between masters and apprentices (Caracausi, 2016), which might justify different treatment in terms of salary and other conditions.

Several other socio-economic aspects of venetian apprenticeship will be investigated, such as: 1) social networks, especially in their social and geographical dimentions. Of particular importance in this respect is the role of guarantors as social brokers. 2) The investigation of possible prototypical careers in different crafts. We presented here just an example of the multifaceted perspectives which could be used to investigate the *Garzoni* database. Other disciplines, such as the history of art and social history, are especially involved. As more and more data will become available, a positive data-driven cycle of historiographical investigation can clearly take shape.

9. Policy and Institutional framework

Co-funded by the Swiss National Science Foundation (SNF) and the French National Research Agency (ANR), the *Garzoni* project started in 2015 for a duration of three years. It gathers three academic partners in an interdisciplinary team which comprises historians, historians of art and computers scientists from, respectively, the history departments of Rouen[12] and Lille[13] Universities and from the Digital Humanities Laboratory of the Swiss Federal Institute of Technology in Lausanne[14]. Closely associated institutions are the Venetian State Archive[15] and the University Ca'Foscari of Venice[16].

With respect to data policy, we plan to publish the ontology, the source code of the tools and the acquired dataset under a CC BY-SA license (Attribution-ShareAlike), while the scans of the primary sources will - hopefully-, be published under a CC BY-NC-ND (Attribution-NonCommercial-NoDerivs).

Dissemination of the work is mainly made through a dedicated blog[17], hosted by *hypotheses*, a platform for academic blogs in the humanities and social sciences. This blog gathers information about the development of the project with information on, among others, historical background, organisation of events, or where to find the data.

10. Conclusion

We have presented an overview of the first milestones of the *Garzoni* project, mainly dedicated to the design and development of tools to extract and render information contained in the *Accordi dei Garzoni*. This overview illustrated the fruitful collaboration between partners of different disciplines to bring forward a digital project. Computational methods and algorithms are truly challenged by 'new' material coming from the humanities (e.g. with modeling issues and record linkage) and historians will benefit from new ways of exploring archival material. Future work consist in proceeding with the project's plans with, among others, continuation and improvement of the already developed tools, annotation evaluation, profession taxonomy building, geo-linking and map-based visualization, data-set interlinking and final publication, study of complementary sources and, naturally, historical investigations.

---

[12] Groupe de Recherche d'Histoire (GRHis, http://grhis.univ-rouen.fr/grhis/)

[13] Institut de Recherches Historiques du Septentrion (IRHiS, http://irhis.recherche.univ-lille3.fr/)

[14] EPFL DHLAB (http://dhlab.epfl.ch/)

[15] http://www.archiviodistatovenezia.it/siasve/cgi-bin/pagina.pl?Lingua=en

[16] http://www.unive.it/pag/13526/

[17] http://garzoni.hypotheses.org/

Acknowledgements

Authors gratefully acknowledge the support of the FNS/ANR grant No. CR12I1L_156272.

References

[Alfani 2011] Alfani, G. Il Grand Tour dei Cavalieri dell'Apocalisse. Marsilio, (2010).

[Artiles 2008] Artiles, J. Sekine, S., Gonzalo, J., 2008. Web people search: results of the first evaluation and the plan for the second. Proceedings of the 17th international conference on World Wide Web, ACM, (2008).

[Auer 2011] Auer, S., Lehmann, J., Ngomo, A. N., 2011. Introduction to linked data and its lifecycle on the web. In ReasoningWeb. Se- mantic Technologies for theWeb of Data, pages 1–75. Springer, (2011).

[Bloothooft 2015] Bloothooft, G., Christen, P., Mandemakers, K., & Schraagen, M. Population Reconstruction. Springer, (2015).

[Bellavitis 2006] Bellavitis, A. Apprentissages masculins, apprentissages féminins à Venise au XVIe siècle. *Histoire urbaine*, 15 (1), pp. 49-73, (2006).

[Caracausi 2016], The price of an apprentice : contracts and trials in the woollen industry in sixteenth century Italy, *Mélanges de l'Ecole Française de Rome, Italie-Méditerranée*, 128-1, *Familles laborieuses. Rémunération, transmission et apprentissage dans les ateliers familiaux de la fin du Moyen Âge à l'époque contemporaine en Europe*, ed. by A. Bellavitis, M. Martini, R. Sarti, (2016).

[Colavizza 2016] Colavizza, G., 2016 (forthcoming). A View on Venetian Apprenticeship through the Garzoni Database. In: Garzoni: Apprendistato, Lavoro e Società a Venezia e in Europa, XVI-XVIII secolo, Venice, Italy, October 10-11 (2014).

[Colavizza et al. 2016] Colavizza, G., Ehrmann and M., Rochat, Y., 2016. A Method for Record Linkage with Sparse Historical Data. In: Digital Humanities Conference 2016, Krakow, Poland, July 11-16 (2016).

[De Munck et al., 2007] De Munck, B., Kaplan, S. L. , Soly H. (eds), *Learning on the shop floor. Historical perspectives on apprenticeship*, London/New York, (2007).

[Doerr 2011] Doerr, M., Theodoridou, M. CRMdig: A generic digital provenance model for scientific observation. Proceedings of TaPP'11: 3rd, USENIX Workshop on the Theory and Practice of Provenance, (2011).

[Epstein 1998] Epstein, S. R. (1998). Craft Guilds, Apprenticeship, and Technological Change in Preindustrial Europe. *The Journal of Economic History*, 58 (3), pp. 684-713, (1998).

[van Hage 2011] van Hage, W. R., Malaisé, V., Segers, R., Hollink, L., & Schreiber, G. Design and use of the Simple Event Model (SEM). Web Semantics: Science, Services and Agents on the World Wide Web, 9(2), 128–136, (2011).

[Infelise 1997] Infelise, M. La crise de la librairie vénitienne, 1620-1650. In: Barbier, Frédéric et al. (eds.), *Le livre et l'historien: études offertes en l'honneur du professeur Henri Jean-Martin*. Geneva: Droz, pp. 343-352, (1997).

[Janowicz 2014] Janowicz, K., Hitzler, P., Adams, B., Kolas, D., Vardeman II, C. Five Stars of Linked Data Vocabulary Use. Semantic Web, 5-3, p. 173, IOS Press, (2014).

[Ji 2011] Ji, H., Grishman, R. Knowledge base population: Successful approaches and challenges. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, p. 1148, Association for Computational Linguistics, (2011).

[Gadd and Wallis 2006] Gadd, A. Wallis, P. Guilds and associations in Europe 900 – 1900. Centre for Metropolitain History, Institute of Historical Research, London (2006).

[Heath 2011] T. Heath, Bizer, C.. Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology, 1(1):1–136, (2011).

[Lanaro 2008] Lanaro, P. Corporations et confréries: les étrangers et la marché du travail à Venise (XVe-XVIIIe siècles). *Histoire urbaine*, 21 (1), pp. 31-48, (2008).

[Lehmann 2015] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mende, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C. DBpedia – A Large-scale , Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web, 6-2, (2105).

[Minns 2013] Minns, C. and Wallis, P. The price of human capital in a pre-industrial economy: Premiums and apprenticeship contracts in 18th century England. *Explorations in Economic History*, 50, pp. 335-350, (2013).

[Ogilvie 2014] Ogilvie, S. The Economics of Guilds. *The Journal of Economic Perspectives*. 28 (4), pp. 169-192, (2014).

[Ogilvie 2011] Ogilvie, S. Institutions and European trades, Merchants Guilds 1000-1800. Cambridge University Press. (2011).

[Rapp 1976] Rapp, R. T. Industry and Economic Change in Seventeenth Century Venice, (1976).

[Sella 1997] Sella, D. Italy in the Seventeenth Century, Routledge, (1997).

[Shen 2015] Shen, W., Wang, J.,Han, J. Entity linking with a knowledge base: Issues, techniques, and solutions. Transactions on Knowledge and Data Engineering, IEEE, volume 27-2, p443 (2015).

[Vrandecic 2014] Vrandecic, D. and Krötzsch, M. Wikidata: A free collaborative knowledgebase. Commun. ACM, 57(10):78–85.

[Wallis 2010] Wallis, P., Webb, C., Minns, C. Leaving Home and Entering Service: The Age of Apprenticeship in Early Modern London. *Continuity and Change*, 25 (3), pp. 377–404 (2010).