frontiers in Digital Humanities

ORIGINAL RESEARCH
published: 11 March 2019
doi: 10.3389/fdigh.2019.00004

Index-Driven Digitization and Indexation of Historical Archives

Giovanni Colavizza[1,2,3]*, *Maud Ehrmann*[1]* and *Fabio Bortoluzzi*[4]

[1] Digital Humanities Laboratory, École Polytechnique Fédérale de Lausanne, The Alan Turing Institute, Lausanne, Switzerland, [2] The Alan Turing Institute, London, United Kingdom, [3] Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, Netherlands, [4] Archivio di Stato di Vicenza, Vicenza, Italy

The promise of digitization of historical archives lies in their indexation at the level of contents. Unfortunately, this kind of indexation does not scale, if done manually. In this article we present a method to bootstrap the deployment of a content-based information system for digitized historical archives, relying on historical indexing tools. Commonly prepared to search within homogeneous records when the archive was still current, such indexes were as widespread as they were disconnected, that is to say situated in the very records they were meant to index. We first present a conceptual model to describe and manipulate historical indexing tools. We then introduce a methodological framework for their use in order to guide digitization campaigns and index digitized historical records. Finally, we exemplify the approach with a case study on the indexation system of the *X Savi alle Decime in Rialto*, a Venetian magistracy in charge for the exaction—and related record keeping—of a tax on real estate in early modern Venice.

Keywords: archives, digitization, digital archives, indexation, information retrieval, Venice, Republic of Venice

OPEN ACCESS

Edited by:
Jeannette Franziska Frey, Bibliothèque Cantonale et Universitaire - Lausanne, Switzerland

Reviewed by:
Daniel Lopresti, Lehigh University, United States
Federico Nanni, Universität Mannheim, Germany

***Correspondence:**
Giovanni Colavizza
gcolavizza@turing.ac.uk
Maud Ehrmann
maud.ehrmann@epfl.ch

Specialty section:
This article was submitted to Cultural Heritage Digitization, a section of the journal Frontiers in Digital Humanities

Received: *02 February 2018*
Accepted: *11 February 2019*
Published: *11 March 2019*

Citation:
Colavizza G, Ehrmann M and Bortoluzzi F (2019) Index-Driven Digitization and Indexation of Historical Archives. Front. Digit. Humanit. 6:4. doi: 10.3389/fdigh.2019.00004

1. INTRODUCTION

Digitization efforts are slowly but steadily contributing an increasing amount of facsimiles of cultural heritage documents. Initiated in the 1980s with small scale, in-house projects, the "rise of digitization" grew further as the World Wide Web developed in the 1990s until reaching, already in the early 2000s, a certain maturity with digital repositories fueled by large-scale, industrial-level digitization campaigns (Terras, 2011). Overall, the successful completion of many projects, led by both private and public sectors, has fostered the definition and adoption of standards and best practices, enabled a better understanding of the costs—by now more predictable—and allowed to earn a good experience to draw upon (Lynch, 2002; Tanner, 2006). As a result, it is nowadays commonplace for many memory institutions to create and manage digital repositories which, among other core benefits, offer rapid, time- and location-independent access to documents (or surrogates thereof), allow to virtually bring together disperse collections, and ensure the preservation of fragile documents thanks to on-line consultation (Deegan and Tanner, 2002; Rikowski, 2008). Importantly, beyond the preservation of and the access to documents, the availability of digital cultural resources also bears the potential of new forms of digital scholarship, communication and education (Boonstra et al., 2004; Meroño-Peñuela et al., 2015).

However, despite this significant momentum, cultural heritage digitization still faces several challenges in enlarging its scope, strengthening its methods and increasing its added value. Quantitatively speaking first, digitization campaigns have only touched the tip of the iceberg, at least in Europe and particularly for archives. The 2015 EU digitization survey of about 1000 cultural heritage institutions reports that on average 23% of European collections have been digitized so far,

with archives having the smallest share (13%), behind libraries (19%), and museums (31%) (Nauta and van den Heuvel, 2015). Digital collections are thus growing, but still represent a modest fraction of cultural holdings.

Next, digitization encompasses several complex processes and institutions embarking on digital projects need well-thought strategies. Besides long term digital preservation (Evens and Hauttekeete, 2011) and licensing matters (Terras, 2015b), one of the key issues institutions need to deal with is the selection of material. *Where to start?* and *Shall we digitize everything?* are the very first questions when facing kilometers of shelving with a digitization objective in mind. Myriads of guidelines exist to answer these questions, exposing various criteria such as legal issues, stakeholders concerns, known use of collections, purpose of the digitization and, naturally, cost and physical condition of documents (Lopatin, 2006). Overall, if surveys reveal important disparities among practices, the selection of materials first and foremost depends on the general context within which digitization takes place, and follows a mix of preservation, user and/or exploitation-driven approaches (Ooghe et al., 2009). According to the EU study, only a third of the surveyed institutions have a written digitization strategy, but this lot is on the rise. This is fortunate, for having a clear digitization strategy has proven to be strongly correlated with the amount of digitization carried out (Borowiecki and Navarrete, 2017).

Finally, if documents eventually become accessible through digital repositories, their retrieval is almost exclusively based on contextual information (*metadata*), and requires, most of the time, the complementary expertise of archivists as well as a thorough knowledge of source holders' history, function and activities (Evans, 2007). In practice, this means that a scholar uses the same means of searching whether visiting a physical institution or a digital repository. In order to improve access to and use of digital historical records, not only documents and their metadata need to be processed and made accessible, but also and mainly their *contents*. As anticipated by Lynch (2002), extracting and linking the complex information enclosed in digitized collections represents the next and natural challenge brought about by digitization.

From an information technology point of view, processing the contents of documents usually falls within one of the two broad families of applications, namely information retrieval and information extraction[1]. Information retrieval (IR) corresponds to the activity of retrieving a specific *document* within a collection or, specifically, to the activity of "finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" (Manning et al., 2008, p. 1). IR systems rely first and foremost on indexation, that is to say the process of building indexes or list of terms present in documents with pointers to the pages in which they occur. In the context of contemporary, born-digital documents, IR capacities enable users to retrieve an item — on the web, an intra-net or a personal computer—thanks to the extensive indexation of resource contents mixed with other means of (hyper)link identification (Brin and Page, 2012). Information extraction (IE), in turn, corresponds to the activity of finding a specific *information* within large volumes of texts or, specifically, to the activity of "[creating] a structured representation (such as a data base) of selected information drawn from texts" (Grishman, 1997, p. 1). IE systems rely on the extraction of salient facts about predefined types of entities, events and relationships in free texts, and on the structured representation of the information extracted thereof (Piskorski and Yangarber, 2013). IE capacities are essential for the construction of knowledge graphs and support a variety of data mining-based applications.

With historical records made digital, indexation of historical contents and extraction of structured information become possible. Considering contents from a lexical perspective, indexation needs to focus on words and multi-word expressions, requires full transcription and will allow full-text search. Considering contents from a referential perspective instead, indexation needs to focus on entities (e.g., persons, locations, etc.), requires partial transcription and semantic tagging, and will allow semantic search. This will lead to increased exploration capacities, with the possibility to not only browse inventories but also to search, retrieve and discover documents and entities of interest across collections, in both IR and IE fashion. If large-scale indexation and semantic annotation are feasible on printed material, to the point where OCR technologies have naturally determined what to digitize (Putnam, 2016), how feasible are they on older, handwritten archival records? And, more generally, *how to turn archives into information systems?*

This challenging and open task requires to consider several factors with, among others, the complex and slow-pace digitization of archival documents, the high cost of their manual indexation beyond metadata (Evans, 2007), and their rich and complex information structure. In a context where no perfect solution exists, neither in terms of digitization prioritization, nor in terms of indexation costs and information relevance, we propose an approach that reconciles those factors in view of scaling indexation efforts and bootstrapping the creation of content-based archival information systems. At its core is the usage of *historical (archival) indexes* as a way to streamline both digitization and indexation processes. This approach stems from the conviction that the real promise of digitized historical archives is to provide quick and complete access to the contents of records, i.e., to index and make them searchable, and that there is a need to scale indexation efforts in order to avoid simply duplicating the archives in the digital space, or to create ever-increasing indexation backlogs.

The remaining of this paper is organized as follows. Section 2 introduces the index-driven approach to the digitization and indexation of archives. Section 3 details the methodology, with a description of the conceptual model and the procedure. Section 4 exemplifies its application on a case study at the Venice State

[1]These concepts grew out of research fields (Natural Language Processing, Information and Communication Technology) dealing almost exclusively with contemporary textual documents which, for a major part, are "simpler" to process—at least for the first steps—than cultural heritage ones. Nevertheless, these concepts remain fully valid and offer a strong base to guide developments around cultural heritage documents.

Archive. Section 5 summarizes the related work and, finally, section 6 discusses the approach and concludes.

2. APPROACH

Archives are not simple collections of documents, but aggregates created by persons, families or organizations within the scope of their affairs, and organized accordingly. The organization of an archive allows the selection of some records of interest, relying on contextual information, and then to search more specifically via browsing. In order to search an archive and filter the documentation to browse for specific information, it is therefore important to appreciate the context and logic of its production. The two core tenets of modern archival science, the principles of original order and of provenance, reflect this approach to research, which relies on an historical appreciation of the entity which created the archive, and of the archive's original ordering and internal logic.

In has been argued that born digital archives are less constrained in this respect. With born-digital archives, this approach to search is no longer the only option, and thus becomes of less importance than for their analog counterparts (Bailey, 2013). The same would apply to digitized archives: when different series and record groups are interrelated through the indexation of their contents, the institutional context and original ordering become just one of several ways to search the archive. As stated by Yeo (2012, p. 71) the "logical associations of records extend beyond the records themselves and embrace relations with other entities in the wider world" and multiple virtual organizations of archival records can therefore be built on top of a single physical organization.

In fact, the context-driven approach to archival research, embodied in the finding tool of archival inventories, reflects but one possible way to search aggregates of documents for specific information. Alternatives require, instead, a shift toward a content-driven approach (Moss et al., 2018). As archives become digitized, the issues of transparent and effective access become increasingly important. Born digital and digitized archives are indeed fostering more opportunities and demands for content-driven approaches: search by keywords, indexed content, as well as means to cope with the lack of reliable OCR, are all highly requested features of digital and digitized archives (DeRidder and Matheny, 2014).

We identify four **indexation stages** of archival material which correspond to different *types of information management* offering different *affordances* in terms of search and finding tools. These approaches are not mutually exclusive but represent instead complementary ways to archival search:

1. **Metadata-based indexation**, for *contextual information retrieval*: this approach relies on the context of creation of records, and ideally on their physical arrangement reflecting it. The main finding tool is the archival inventory. This is the mainstream approach currently followed to build information systems for historical archives, and its affordance is browsing, i.e., a set of archival units is selected via contextual metadata and then perused thoroughly in search for specific information.
2. **Entity-based indexation**, for *referential information retrieval*: this approach relies on selected searchable contents (entities of interest), for the purpose of finding possibly relevant documents. The main finding tool is the index, linking entities with relevant documents (where they are mentioned). Its affordance is a mix of searching (for entities) and browsing (documents), typically on a much smaller search space than using the contextual metadata approach. The referential IR approach is in use today as it was historically.
3. **Structured entity-based indexation**, for *information extraction*: this approach still focuses on selected contents (entities of interest), but entails the extraction and creation of structured information from the relevant documentation. Its affordance is searching, and the main finding tools are databases. The IE approach is in use today as it was at times historically.
4. **Full content-based indexation**, for *information retrieval*: these approaches rely on the full availability of contents, thus are by far and large only possible in a digital setting. The affordance is searching, and the finding tool is the (full-text) search engine. Google Books is an example of an IR approach to search, on a collection of digitized objects.

In this article, we propose and formalize an approach to jointly bootstrap the creation of information systems for historical archives and guide their digitization, following a referential information retrieval approach (number 2 above). More precisely, we propose to leverage *historical indexes* to, on the one hand, extract their index data and thereby rapidly feed a digital information system and, on the other, provide a way to prioritize digitization, focusing on record aggregates with rich indexation systems. In fact, several aggregates of records in historical archives were produced by public institutions which, at times, enriched them with simple or advanced indexation systems, in a purely referential IR fashion. The purpose of these indexation systems was the same as modern ones: allowing users to find all the relevant information pertaining to a certain entity of interest. As a consequence, historical archival indexes commonly focused on indexing things such as persons, events and topics, interlinking them to their relevant records by alphanumerical references (the equivalent of "foreign keys" in database terminology).

The Oxford Dictionary defines an index as "an alphabetical list of names, subjects, etc. with reference to the pages on which they are mentioned[2]." An *archival index* is, similarly, a tool to recover information regarding relevant entities or topics across records. In the context of archives, indexes can be found at the level of individual archival units, say to index the contents of a register (like the index of a book), or at the level of larger aggregates, in which cases indexes themselves are compiled into individual registers. Even more complex indexation systems can sometimes be found, usually helping to access a set of related document series within an archive. Our approach entails the

[2]http://www.oxforddictionaries.com/definition/english/index

alignment of different indexes into a unique meta-index: a superior layer of indexation, which subsumes all the information from individual indexes and integrates it into a global referential information system.

The proposed approach has multiple benefits. From an archival view point, it bootstraps a referential access to contents intrinsically informed by the original ordering of the archive, since the index data was (usually) produced when the archive was still current. At the level of the construction of an information system, historical indexes provide a first set of index data easily expandable thereafter, which can then be used to index previously non-indexed records too. With regard to automation opportunities, indexes are usually more amenable to semi of fully automatic processing because of their more regular layout and handwriting. As per digitization strategy and planning, indexes provide a way to guide efforts, by focusing on indexes and well-indexed record groups. And finally, content-driven access might help widening the interested user base, as it provides a lower barrier to entry than a contextual approach. Our work in this respect is in line with the efforts to build entity-centric information retrieval systems for historical contents (Boschetti et al., 2014; Coll Ardanuy et al., 2016; Menini et al., 2017). Overall, this methodology takes into account both the long-term goal of content-driven access and the short-term constraints of resources (costs) and technology (current limits to automation, especially OCR of handwritten documents).

We adopt in this article the archival terminology defined in Theimer (2012), along the lines of the archival terminology curated by the International Council on Archives[3]. An *archive* is a collection of materials created or otherwise accumulated by an entity, be it a person, a family, an organization, in the conduct of its affairs. A second, important specification is that records are typically assembled in aggregates according to their origin, often related to the entity and activity which generated them. We consider a *record group* as the aggregate of records generated by a single entity, and a *document series* as the smaller groups contained therein, which originate from a clearly individuated activity of the same entity. For example, a city archive might contain the vital records group, including birth, marriage and death certificates among its document series.

In the following section we introduce a methodology to work with indexes and to use them in a digitization campaign in view of a content-based information system.

3. METHODOLOGY

Using historical indexes and the data they contain in a modern information system calls for a methodology that takes into account both the archival and the computational dimensions. From an archival perspective, indexes needs to be appreciated as historical documentation with a specific archival context, purpose and history. From an information system perspective, they correspond to a source of index data, whose quantity and quality need to be assessed. This requires the joint expertise of computer scientists and archivists.

[3]http://www.ciscra.org/mat

Firstly, a conceptual model of historical indexes needs to be developed. This model should be generic and flexible enough in order to, on the one hand, allow its wide application and the procedural alignment of archival indexes and, on the other, its adaptation to the varied nature of historical records. Secondly, a data assessment and acquisition procedure takes place. In this regard, indexes need to be described and compared with respect to the conceptual framework, and their quality needs to be assessed for the indexation purpose at hand. Such description and comparison entail the consideration of the indexed entity typology, the quantity and quality of index data, the coverage of the index and its relation with other indexes within the same archive. These processes require, all along, the appreciation of indexes as historical documentation. Data acquisition can then be done via the manual or automatic extraction, semantic annotation and alignment of index contents via entity and record linkage. The focus of this paper is not on the technicalities of data acquisition and alignment.

We begin by defining a set of concepts to describe the constituent elements of historical indexes, which can support their formal description and comparison. We then outline the series of steps to be taken in order to work with historical indexes in view of building an information retrieval system.

3.1. Conceptual Model

The first step in order to work with historical inexes as a sourc of index data is to define a conceptual model to describe and represent the information they contain. To this end, we introduce a set of definitions as a conceptual model (the following reads better when looking at **Figure 2** at the same time):

- *Thing*: an entity or a concept of interest, which is mentioned once or more in a set of records and is used to index them. Examples of things of interest are persons, places, dates, topics. Entities and concepts are common entry-points not only of indexes, but also of modern information systems and wiki-like systems.
- *Indexation unit*: a mention of a thing in an index. Mentions can be quite elaborated. For example, in the case of person mentions, several components could be used to refer to a person: name, surname, patronymic, origin, profession, family relations, and others. The set of mention components and their order, usually relatively uniform in an index, we call *naming convention*. We consider the whole mention as a single indexation unit, and abstract from a detailed appreciation of its components at this level. Naming conventions will become relevant for the task of aligning several indexes into a unique meta-index.
- *Information unit*: an optional set of information which further specifies an indexation unit. Examples are chronological spans which could indicate the period of validity of the documentary evidence referred by the entry of the index.
- *Index reference*: an identifier which redirects an entry of an index to (the identifier of) a record or a piece of further documentary evidence/information. References are usually alphanumerical, often incremental numbers such as page or sheet numbers of different registers.

- *Entry of an index*: at the very minimum, an indexation unit with a reference. More generally, an indexation unit with zero or more information units and one or more index references pertaining to it.
- *Indexed information*: the information, record or other evidence the index points to via references. Normally, one indexed information per index reference exists (thus one or more indexed information per entry/indexation unit).
- *Index*: a (possibly sorted) list of indexation units, which stand for some things which are indexed, and refer to some indexed information via references.
- *Indexation system*: a set of indexes which index a coherent set of document series. An indexation system is for example composed of indexes which allow access to complementary records with information on the same entities or concepts.
- *Meta-index*: the result of the alignment of several indexes into a unique one, by merging indexation units which refer to the same thing, keeping their references to indexed information.

In summary, things such as persons are mentioned, explicitly or implicitly, in multiple records of interest, which generically compose the indexed information. These documents are indexed via mentions of things, mostly for practical reasons, such as rapid access to indexed information. In the index we will often find one entry per thing, e.g., a row if the index is organized into lines. Each entry contains one indexation unit, which is the mention of the associated thing, possibly using a uniform naming convention across all mentions in the index. Index entries also contain one or more references to the indexed information, and possibly some information units to further specify it.

An example of part of an annotated index is given in **Figure 1**. The page is the beginning of an alphabetical index of fiscal persons (things), letter A. Every row is an entry, starting with the indexation unit which is the mention of the person. After a justification line, a set of references is given to items listed in the ensuing pages. For example the first line starts with the indexation unit *Antonio Grimani*, a Venetian nobleman, who is linked with items 1, 104, 105, 106, 157, 158. The second unit is an organization, the *Abbazia della Misericordia*, an abbey, and so forth. A sketch of the components of a historical index, according to the descriptive model just defined, is given in **Figure 2**. Eventually, multiple indexes are aligned into a unique meta-index which constitutes the first bulk of the archival information system. From the meta-index, all indexed records relevant with respect to a thing of interest can be accessed.

The proposed method considers historical indexes as sources of index data. If historical indexes are assessed, described, digitized and transcribed, they can then be aligned into a unique meta-index. The alignment simply consists into the detection of all indexation units from different indexes which refer to the same thing. Having done this, the skeleton of the information system is in place, and further digitization and indexation can proceed at any pace.

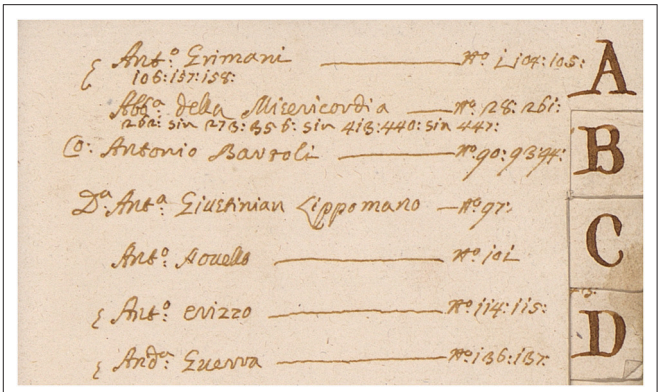

FIGURE 1 | An example of a historical index. There are seven entries organized into lines, each with a mention of a thing (indexation unit, in this case fiscal, or juridical persons) and a list of references (index references). The last entry transcribes as: "Ser Andrea Guerra ... numero 136, 137." From: State Archive of Venice, *X Savi alle decime in Rialto*, *Catastici delle parrocchie*, 1740, *San Marcilian*, f. 436, c. 4v.

3.2. Procedure

Having defined a conceptual model to describe historical indexes, let us detail the procedure. First, an archive undergoing digitization should be surveyed in search for indexes. These indexes should be analyzed, in view of their possible alignment during *phase 1 (archival survey and contextualization)*, whose result is a tentative plan for their digitization. In this preliminary phase, indexes should be thoroughly understood in terms of their normative and material contexts: for example, if mentions of persons are indexed (alphabetical index), it is important to know which persons ought to be (and actually are) included in the index and why. Secondly, a description of the indexes, along with a technical plan for their integration should be produced during *phase 2 (formalization and design of integration)*. In this phase, the selected indexes are compared and described, and an annotation model of their contents is produced by detailing or extending the general ontology. Thirdly, indexes should be digitized, transcribed and annotated in their constituent parts during *phase 3 (acquisition)*, using the annotation model defined at phase two. Lastly, a manual or automated procedure for text normalization and index alignment should build or integrate the meta-index during *phase 4 (alignment)*.

After phase one, a digitization strategy can be defined. The availability and quality of indexes can in fact determine which records should be prioritized for digitization, *caeteris paribus*, starting from indexes and records with consistent indexation systems. It must also be noted that starting from indexes can support the indexation of records without indexation systems of their own, provided they contain mentions of things already indexed.

3.2.1. Phase 1: Archival Survey and Contextualization

This phase entails the critical inspection of the record groups which contain some indexation tool, specifically made in order to probe the archive for indexes, individuating and evaluating:

1. which *things* are indexed;
2. what is the *coverage and quality* of the indexes and of indexed information;

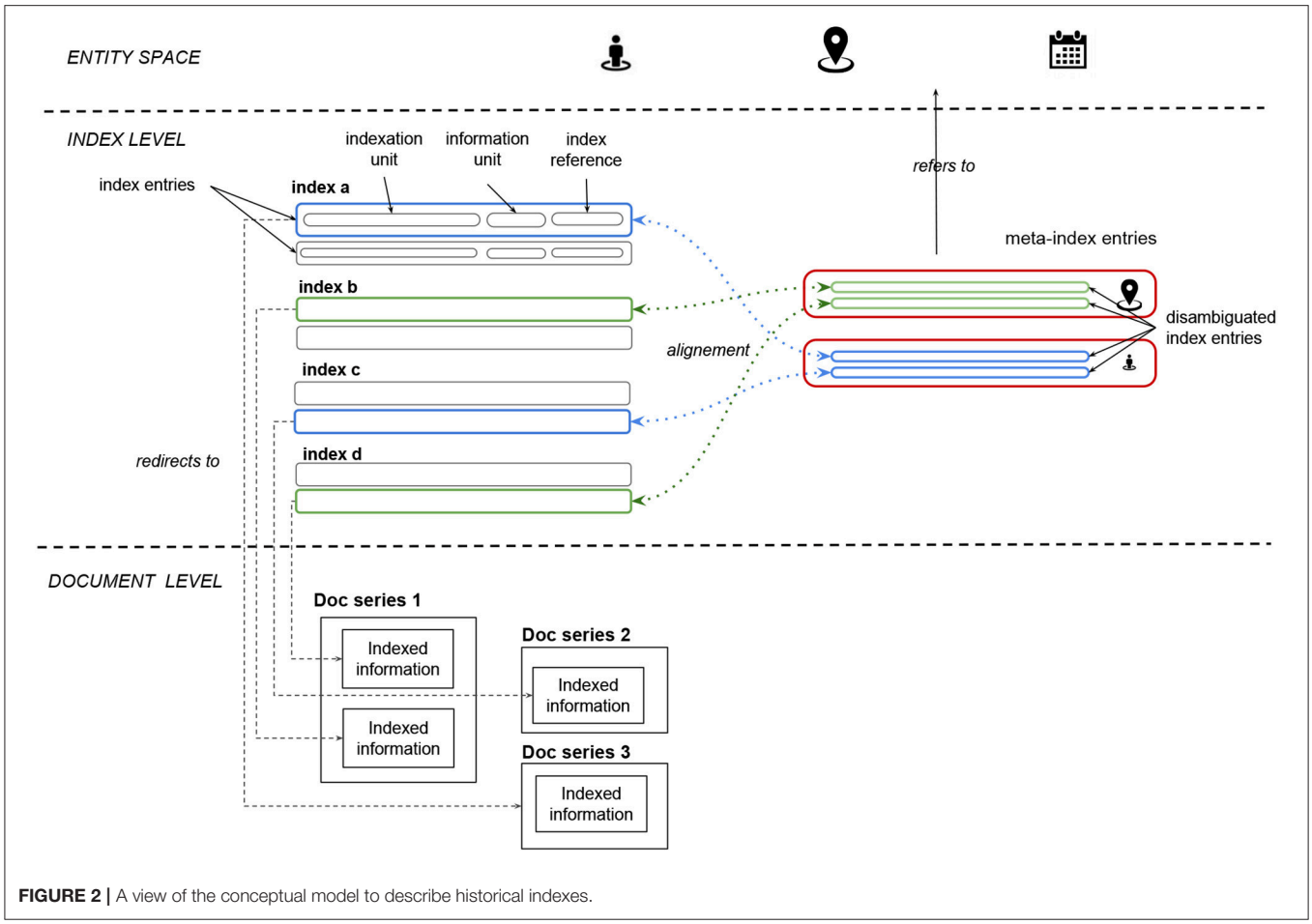

FIGURE 2 | A view of the conceptual model to describe historical indexes.

3. what is the *degree of interrelatedness* of different indexes within the same group of records, which is to say the evaluation of the indexation system, if present;
4. what is the *degree of overlap* of several indexation systems spanning across different record groups, if applicable.

Steps 1-2 should be evaluated for every single index, step 3 is an intermediate and possibly conclusive step, at the level of the whole record group or archive and step 4 is even more generic and might not always apply. Given a thorough execution of these steps, it will already be possible to inform the digitization effort by establishing a priority over records according to the quality of their existing indexation. At the same time it will be possible to design the integration of all the selected indexes into the information system. As much as this decision assumes some knowledge of the presence and quality of the indexation tools for the records at hand, it also depends on the priorities of the project and available resources, whose evaluation is beyond the scope of this article. The process we consider in what follows assumes that a set of record groups containing indexes was selected as a possible candidate for a digitization campaign.

When considering an index, we need to identify and qualify the things being indexed: this should be immediately straightforward. An issue might arise from indexes where more than one type of thing is indexed, such as both topics and persons. In that case, and for what follows, we shall consider all entries for each type of thing as constituting a separate index to work with.

The next step is the evaluation of the coverage and quality of every index, in a quantitative estimation. The coverage of an index provides an idea on how broad the index should be in theory, and actually is in practice. Initially, we consider the normative context of the indexation procedure as a way to define the theoretical coverage of an index, i.e., what the index is supposed to contain. In practice, however, we assess the **real coverage of an index** with two measures:

1. The first measure assesses the coverage of an index with respect to a global population of interest. Concretely, the population coverage measure Cp_i considers the number of indexation units N_i of an index i over a global population of interest N, that is to say the fraction of this population which is indexed, or $Cp_i = N_i/N$. It is to be noted that the choice of the population of interest determines the result of an index coverage calculation, and depends on the available evidence and on the goals of the project. Let us consider the

fictive example of an index of booksellers in Venice for the year 1600. We define N - the global population of interest—as the total estimated population of the city for the year 1600, and we define N_i—i.e., the number of indexation units in our index—as the estimate of booksellers active in Venice in the same year. The coverage of the bookseller index is then given by the estimated fraction of booksellers over the whole Venetian population and is, in this case, likely low. Instead of the whole Venice population, it is also possible to use an estimate of the total number of resellers and merchants as a global population of interest. This would result in a much higher coverage measure for the index. This choice would be pertinent in a project that, for example, deals with Venetian economy. Lastly, an estimate of the theoretical coverage of an index could be used in the numerator in place of the real number of indexation units contained in the index. However, the resulting coverage measure would in this case not account only for surviving records, and thus be of less interest in practice.

2. The second measure assesses the coverage of an index with respect to a time span of interest. Concretely, the time coverage measure Ct_i is the chronological time span covered by the index i, considered in comparison with a real or ideal valid span of interest. We define T_i as the number of months or years for which the index survives (this might be affected by historical events such as missing records), and T as the total time-span of interest. As an example, an index of persons might cover the years 1500 to 1800 in theory (according to norm), but only the records for the last 50 years survive, limiting its coverage over time. Then $Ct_i = T_i/T$, in the example $Ct = 50/300 = 0.17$, a relatively low coverage over time.

Calculating the coverage of an index allows to assess its integrity and to evaluate its relevance regarding the objectives of a given (research) project. If an index has a wide theoretical coverage but only a small fraction of its entries have survived, it is perhaps not of much use in the context of indexation. Similarly, if an index is integral, but only covers a tiny fraction of the population of interest, it is also less useful as a source of indexation data.

The **quality of an index**, another quantitative estimation, focuses instead on its internal consistency from an information retrieval point of view, and is given by its:

1. *Precision* P_i, namely the proportion of index entries which correctly point to their indexed information. Given an index i, define as I its total number of entries, and as I_t the number of entries which correctly point to the relevant indexed information. The precision of the index is $P_i = I_t/I, 0 \leq P_i \leq 1$. The ideal precision is 1.
2. *Recall* R_i, which is the proportion of indexed information correctly indexed by index entries, over the total. Define as J the total number of indexed information, whatever these are, and J_t the number of indexed information which are correctly indexed, that is to say which have a corresponding entry in the index. The recall of the index is given by $R_i = J_t/J, 0 \leq R_i \leq 1$. The ideal recall is also 1.

Both measures can either be evaluated exactly, or estimated over samples of data. Note that both concepts borrow from computer science, especially information retrieval (Manning et al., 2008). Evaluating the quality of an index in this respect is necessary to have a grasp of its internal reliability before using the information it contains to populate a meta-index, and ultimately before even considering to work with the index.

The **degree of overlap** is an estimate of the extent to which different indexes index the same information. Its calculation is based on the population coverage measure, i.e., the proportion of indexed things. More precisely, the degree of overlap between two indexes corresponds to the proportion of things indexed in both indexes over the number of things indexed in the shorter of the two. The degree of overlap is bound between zero (no overlap) and 1 (total overlap), and is measured over two indexes which index the same things (otherwise their overlap is 0). Beyond the measure, it is less obvious to state when a certain degree of interrelatedness is to be aimed for. Take for example two alphabetical indexes of persons: the overlap is maximal if both indexes are expected to index the same persons (minus errors and lacunae), and minimal if they are expected to index completely different persons. However, the degree of overlap should be evaluated in the context of a specific project, as say high overlap could be beneficial if two indexes point to different information for the same persons, or not if the information pointed to is very similar and does not justify processing both indexes. The same consideration applies for the overlap of indexes across different record groups. Note that if two indexes have low or zero overlap, but index the same things, they might profitably complement each other.

The outcome of this first phase is a better appreciation of the indexes available for the record groups under consideration for digitization. Different indexes can also be compared systematically in order to devise a digitization plan, as we will show for our case study (cf. Section 4). Typically, a digitization strategy at this stage entails prioritizing the relevant indexes, as well as the most relevant records they index.

3.2.2. Phase 2: Design of Integration

This phase involves firstly, a description and comparison of the selected indexes, and secondly the definition of an annotation model which will be subsequently used to transcribe and annotate indexes. The description of an index shall aim at formally detailing its components, which were individuated during phase one. For every index, the following information is relevant:

1. Indexed things (e.g., persons).
2. Indexation units (e.g., mentions of persons). A detailed description should include all the components of the indexation unit and their ordering, or the naming conventions.
3. Information units (e.g., the date of the registration of the index entry).

4. Index references (e.g., a numerical identifier). Recall that entries could be multi-referential.
5. Indexed information.

3.2.3. Phase 3: Acquisition

Indexes should next be digitized, transcribed and annotated in their constituent parts, using the annotation model defined during phase two. This third step, acquisition, corresponds to the digitization, transcription and annotation of indexes, and can be done with any means available, manual and automated. It is arguably mechanic, despite being the most resource-involved part of the procedure.

3.2.4. Phase 4: Alignment

A possibly automated procedure for normalization and alignment should integrate the transcribed and annotated indexation units into a unique meta-index, so that the same thing mentioned in multiple indexes and records can be accessed via a unique entry point. Even if in principle it might be useful to capture all the components of every entry of every index of interest into the meta-index, in practice only the indexation units will normally be part of the meta-index as mentions of a thing of interest. The task of performing entity/record linkage or named entity disambiguation on historical texts is an open and challenging research area, beyond the intended scope of this paper (Piotrowski, 2012; Olieman et al., 2017; Rovera et al., 2017).

Lastly, it is worth noting that, for meta-indexation to be possible, indexed things (entities and concepts) should be independent from the indexes, i.e., they should exist and/or have a meaning outside of the indexation systems to be aligned. In case of internally built index objects, the alignment is not possible without an explicit effort to harmonize possibly separated conceptual spaces. An example is given by indexes using different taxonomies to index the same indexed objects: their alignment into a unique meta-index would require not only a surface alignment of indexation units, but the alignment of the concept spaces embedded into the two taxonomies.

At the end of the process, two main results will be acquired:

- Users will be able to search all the information contained in a set of indexes, and navigate from the space of things (meta-index) to indexed information, therefore speeding up the process of finding all the indexed information regarding a given thing.
- The information system will be ready to integrate the digitized copies of the records containing the indexed information.

In order to illustrate the process in practice, we discuss its application on the archive of the *X Savi alle Decime* at the State Archive of Venice in what follows.

4. CASE STUDY

The Republic of Venice was considerably advanced with respect to the organization of its archives. Her republican form of government led to an early abundance of documentation, and the realization by the ruling elites that a well-managed archive is a great source of power and control. It does not come as a surprise that the first known treatise on the management of archives, *De Archivis Liber Singularis* by Baldassarre Bonifacio, was printed in Venice in 1632 (Duchein, 1992, p. 16). De Vivo (2010) discusses the strategies put in place by Venetians to order and index the secret archive, a collection of sensible state documents, mainly originating from the Senate. Three ordering principles were used: grouping documents by institution, by subject matter and chronologically. In this context, the main information retrieval device was the *rubrica*, or an index of topics organized hierarchically (taxonomies). These *rubriche* were used to index single registers up to whole document series, and even to index other *rubriche*.

It must be stressed that indexation is an historical process, changing at specific moments during the history of the Republic in order to cope with new needs and a growing mass of documentation.

The *X Savi alle Decime in Rialto* were ten magistrates mainly in charge of the *decima*, a tax of the tenth part of the estimated value of real-estate in Venice and the *Dogado*, the original territories of Venice. The origins of the existing portion of the archive of the *X Savi* date to the year 1514, when a fire destroyed all previous records. The *decima* was initially calculated via a call for fiscal self-declarations (*condizioni di decima*), to be submitted to the *X Savi* by decree of the Senate and renewed at episodic occasions. Seven so called *redecime* exist, for the years 1514, 1537, 1566, 1581, 1661, 1711 and 1740 (Canal, 1908, pp. 122–125). Every *redecima* entailed a collection of self-declarations in the number of thousands, containing the details of the belongings of any fiscal person (individuals, organizations and legal bodies included). When a *condizione* was deposited, a personal account was opened in the registers of the books of exchanges (*Quaderni dei trasporti*), where the transactions involving that person were to be registered from that day until the next *redecima*, which in turn determined the opening of a new series in the *Quaderni*. The *Quaderni* series thus keeps trace of every declared real-estate transaction in Venice from 1514 to approximately 1808. The *Quaderni* have alphabetical indexes with references to the page number where the indexed person account is: this is the first index to be considered in what follows (*Indici dei Quaderni*). Besides the *Quaderni*, two other document series are of interest: the books of movements (*Giornali dei traslati*), containing the summaries of the act which originated a transaction registered in the *Quaderni*, accessible by date; and the *Catastici*, a wholly different series existing only for the *redecime* of the years 1661, 1711 and 1740. The *Catastici* are house-by-house inspections, carried out for every parish in Venice by the local priest with officials from the *X Savi*, conducted in order to provide information to verify self-declarations (i.e., the *condizioni di decima*). As a consequence, in the *Catastici* it is possible to find, for every real-estate unit in the city, information on the owner, the tenant, the usage of the unit and its rent, as a proxy of the unit's market value. These records possess alphabetical indexes for every parish, containing the names of the owners and pointing to the unique identifier of the owned unit(s). The *Catastici* provide the second index analyzed in what follows (*Catastici* indexes). In what follows we

consider the indexes for the records produced during the last survey (*redecima*), that of the year 1740.

4.1. Description of the Two Indexes

The *Quaderni dei Trasporti* series, relative to the *redecima* of 1740, comprises 5 index and 13 account registers (covering the period between 1740 to 1808). Each entry in the indexes points to one or more sheets in the account registers, where the account of the given person is. An index entry is composed of an indexation unit (fiscal person) and a set of index references (sheet numbers, pointing to the 13 account registers). Note that persons are indexed by first name, then by family name, resulting in two nested indexation levels. A person account is composed of a header (the name of the person) and of two columns, in typical double-entry format, registering all transactions which led to a variation in the tax to be paid. An example of index can be seen in **Figure 3**, while examples of indexed information are shown in **Figure 4**.

The *Catastici* series comprises for its part 75 small registers, one for each parish in Venice plus the Jewish *ghetto*, each comprising an index named "*rubriche*." In those *rubriche*, each entry is composed of an indexation unit (fiscal person), plus a set of index references (entry numbers) pointing to the list of housing unit records in the remaining part of the register. Housing unit records comprise five information (the register presents five columns): the record number (incremental from 1, starting anew for each parish), the typology of good, the names of the tenant and of the landlord, and the paid rent. An example of index can be seen in **Figure 5**, while examples of indexed information (housing unit records) are shown in **Figure 6**.

Record keeping processes for *Quaderni* and *Catastici* were different. It is likely that an entry in the *Quaderni* followed either the registration of a new declaration (*condizione di decima*) or the need to continue an already existing account in another part (i.e., another page and/or register), for space was insufficient to record further transactions. In both cases, the index was likely updated at the same time. The *Catastici* were instead small registers, meant to be carried and filled during field inspections. The index was, in all likelihood, prepared only after the inspection of a parish was over. Only afterwards—it is not clear when and by whom—the two series of *Quaderni* and *Catastici* were eventually cross-checked for declaration inconsistencies. Evidently, it is only relying on the two indexes that this last operation could be carried out.

4.2. Archival Survey and Contextualization

The first step of the archival survey and contextualization phase consists in individuating what is indexed. In our case, the *Quaderni* registers contain records of real estate transactions between legal entities (i.e., individuals, organizations or others) owning lands or buildings located in the city of Venice or in the *Dogado*, or elsewhere in the domains of Venice but registered in the city. As for the *Catastici* registers, they contain records of the exact same things as in the *Quaderni*, that is to say landlords and their properties, but only those located in the city of Venice proper (not in the *Dogado* nor other areas outside of Venice)[4].

Next, we need to calculate the coverage of the indexes of these registers, normally by getting an estimate of the proportion of a relevant population which is indexed. In our case, we know the exact numbers:

- The indexes of the *Quaderni* contain 32,406 individual entries (indexation units). The chronological span is continuous from 1740 to 1808.
- The *rubriche* of the *Catastici* contain 12,607 entries. In this case, the real number is actually lower, since there is an index per parish, thus any landlord with estates in different parishes is indexed multiple times. The chronological span is limited to the year 1740.

We can define two possible populations for comparison: an estimate of the total population of Venice, and the very population of landlords. With respect to the latter, $Cp = 1$ for both indexes. With respect to the former, we can take an estimate of the Venetian population for the *Catastici*, and the same but adjusting for time using natality rates for the *Quaderni*. Note that the population of physical persons is but a lower-bound approximation of the population of juridical persons. According to Beltrami (1954, p. 59), the population of Venice was approximately of 149,000 in the year 1760 and of 138,000 in 1696. We can extrapolate approximately 145,000 for the year 1740. Over the years, the population grew (newborns and baptized) by 23,700 units by 1789, and approximately 31,284 by 1806 (Beltrami, 1954, p. 140)[5]. Therefore, we can estimate 176,284 (145,000 + 31,284) persons active in Venice from 1740 to 1806. This figure is obviously a blunt lower-bound, given the limitations of documentary evidence and the disregard for immigration, an important source of population growth. Furthermore, indexed things can be organizations (e.g., religious institutions) and not just real physical persons. Nevertheless, from these estimates we can conclude that the coverage of the *Quaderni* is approximately 32,406 / 176,284 = 18.4%, and the coverage of the *Catastici* is approximately 12,607 / 145,000 = 8.7%. This estimate must be taken as a guideline more than as an exact measure. Its role is to help appreciate the coverage of the indexes with respect to possible populations of interest[6]. With respect instead to the coverage over time, the two indexes differ greatly, as one index spans 69 years, while the other only the year 1740.

We focus next the quality of the indexes. Precision and recall have been estimated on 200 randomly chosen entries per index, which have been manually verified. The *Quaderni* indexes have a $P = R = 0.995$, with only 1 error each. The *rubriche dei Catastici* fare slightly worse, with $P = 0.97$ (6 errors) and $R = 0.96$ (8 errors). The results, albeit taken from a very small sample, are of

[4]More details are in the decree of October 26th 1507 in State Archive of Venice, *Dieci savi alle decime in Rialto*, reg. 2, c. 36v-37r and in Canal (1908, pp. 8–12).

[5]The figure for the last 16 years is estimated using the mean growth of the preceding 5 decades.

[6]In absolute, if independent sources of information are available and can be used to confirm the likelihood the computed numbers are within reasonable bounds, they should be considered. This availability can however not be assumed for all series.

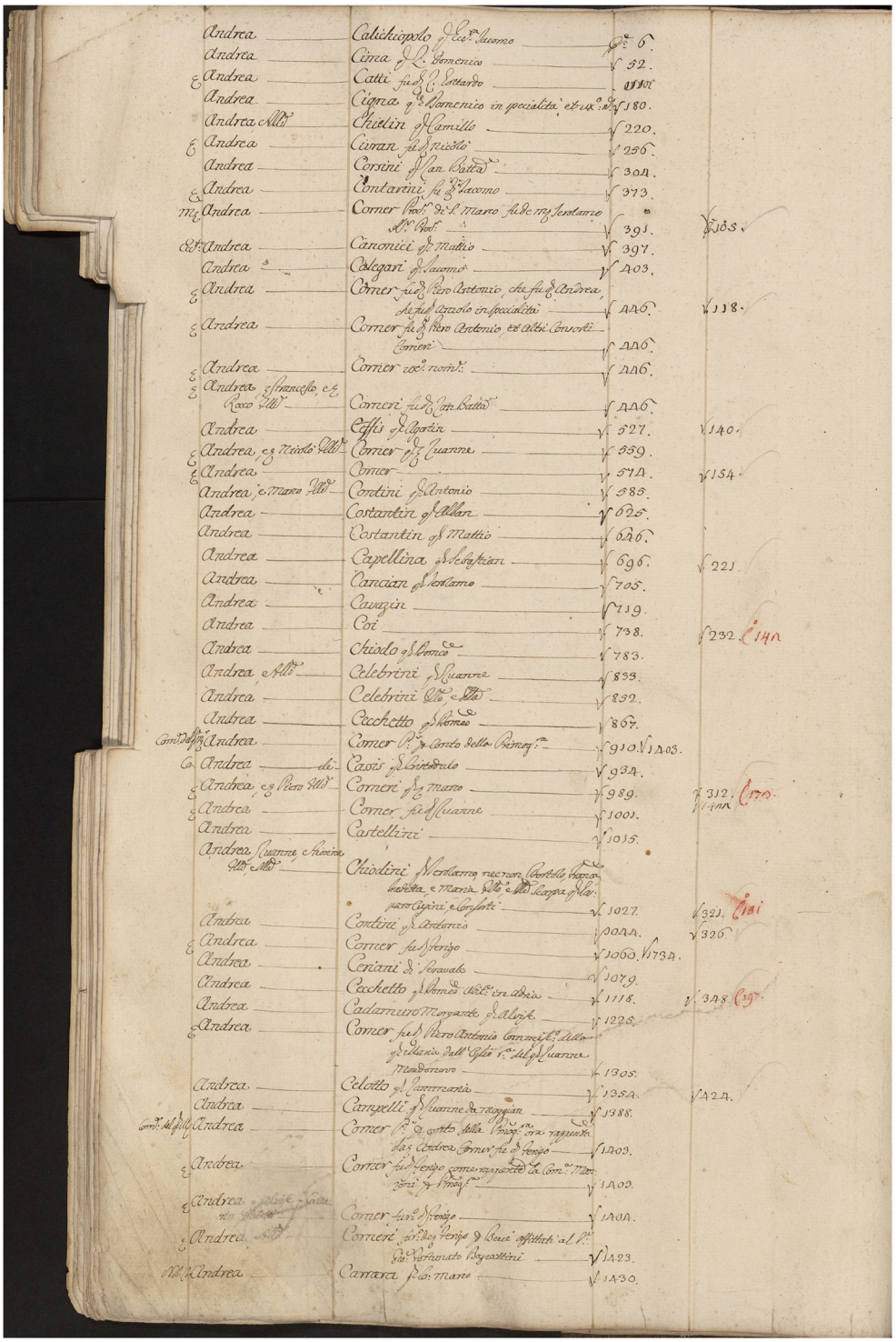

FIGURE 3 | A set of entries in an index of the *Quaderni*. A row just after the middle, 4th row after the page cut refers to "ser Andrea Corner fu domino Zuanne," who has one reference, 1001. From: State Archive of Venice, *X Savi alle decime in Rialto*, *Quaderni dei Trasporti*, r. 1520, c. 32v.

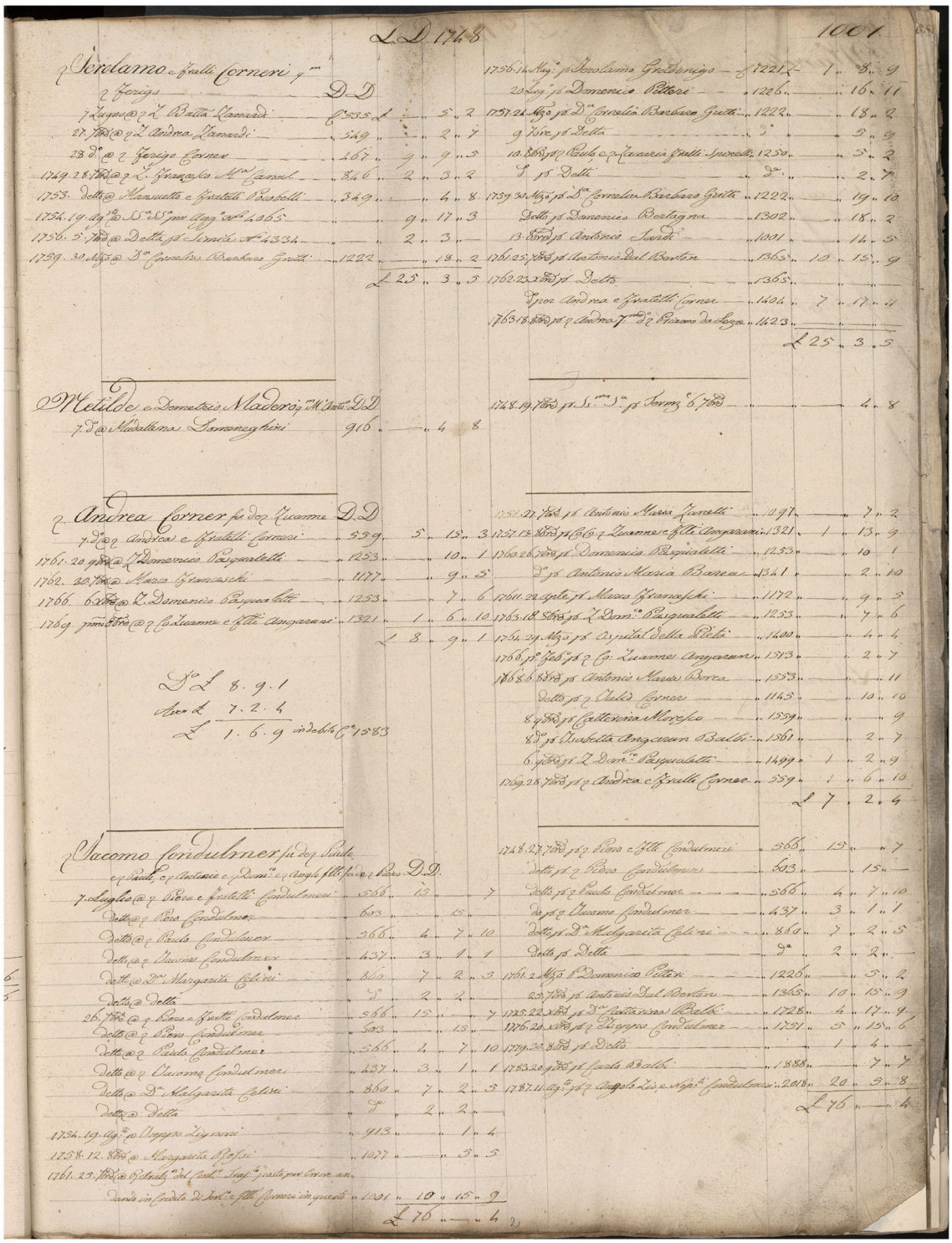

FIGURE 4 | A set of accounts in a register of the *Quaderni*. Note that the sheet is marked as 1001 (top-right corner). The third entry refers to "ser Andrea Corner fu domino Zuanne." There are five accounts per page. Andrea's is filled with transactions. The left column is for newly acquired possessions (for which he needs to give, or *dare*, a certain amount of tax), the right column is for lost possessions (for which he needs to be relieved from a certain amount of tax, or *avere*). From: State Archive of Venice, *X Savi alle decime in Rialto*, *Catastici delle parrocchie*, r. 1529, c. 2r.

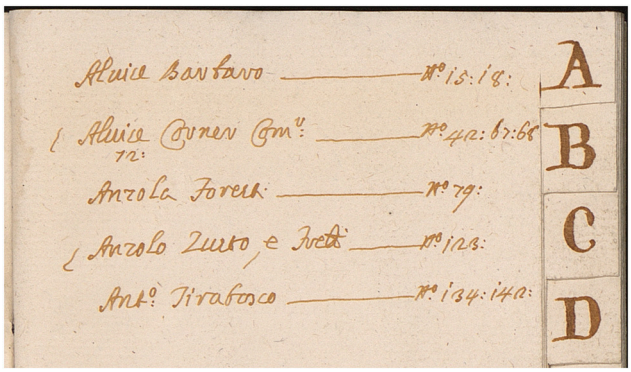

FIGURE 5 | A set of entries in an index of the *Catastici*. The first row refers to "Alvise Barbaro" who has two references, 15 and 18. From: State Archive of Venice, *X Savi alle decime in Rialto*, *Catastici delle parrocchie*, 1740, *San Maurizio*, f. 434, c. 851r.

very high quality and allow to consider the indexes as source of indexation data.

The quality of the indexed information could be verified using a multiplicity of sources. The most important information is related to the possession or exchange of real estate assets recorded in the *Quaderni*, which can be triangulated using notary records or testaments. Another, less difficult verification could rely on the *condizioni di decima*, the self-declarations originating almost every record in the *Quaderni* registers. The usage of real estate, which is investigated in the *Catastici*, could as well be verified using the notarial series. Finally, the degree of overlap of the two indexes is 100%, given that all things indexed in the *Catastici* ought in theory be indexed in the *Quaderni*, even if not necessarily vice-versa.

Table 1 summarizes the main elements of comparison resulting from the first phase. The main concern is the misalignment of the coverage with respect to time, as the *Catastici* provide just a snapshot of a situation, whilst the *Quaderni*, used to track changes, contribute a vision over the years.

4.3. Formalization and Design of Integration, Acquisition, and Alignment

During the second phase we formalize the description of indexes. We learned from phase one that both types of indexes, of the *Quaderni* and of the *Catastici*, consider fiscal persons as main entities of interest. Furthermore, the following elements should be defined with respect to our indexes:

1. *Indexation units* correspond to mentions of fiscal persons. In both cases, person mentions can be more or less refined in a variety of ways. Person names consist of a forename (the main person name in Italian), usually followed by a (male) patronymic. Optionally, titles are added at the beginning and/or the end of a mention, often in abbreviated form. Specific cases such as organizations or juridical bodies can have a more elaborated names.
2. *Information unit* is absent, i.e., no further information specify the indexation units.
3. *Index references* consist of sheet or entry numbers, in Arabic numerals.
4. Finally, the *indexed information* correspond to a double-entry account for the *Quaderni*, and to a housing unit record in the *Catastici* registers.

The model defined for the annotation of the indexes follows this index structure (index entry, mention, reference and indexed information).

After formalization, acquisition can take place (phase three). In this regard, the digitization of indexes was done on site in Venice, as well as their transcription and annotation. The latter was carried out using a web-based transcription and annotation interface which, as its core, is fully compliant with the International Image Interoperability framework (IIIF[7], both image and presentation APIs) and the Web Annotation Data Model[8] standards.

Upon full acquisition of index data, the last phase (four) corresponds to integration, that is to say the normalization of indexation units and their alignment in order to build a meta-index. Here a record linkage approach will need to be developed, perhaps on the model of a previous method developed for another Venetian document series, which combines various similarity measures based on (sparse) context information and person attributes (Colavizza et al., 2016). We leave this as future work.

Finally, in parallel to these acquisition and integration phases, it should be noted that digitization proceeds with indexed documents, i.e., *Quaderni* and *Catastici* registers, and with related document series, such as the *Giornali dei traslati*, and the indexes of other *redecime*.

5. RELATED LITERATURE

This section summarizes previous indexation methodologies and efforts, from the birth of indexes as archival finding tool toward more recent digital approaches. It further discusses the impact of digitization on the work of historians.

Historical indexes were for the most part produced with the goal of indexing specific document series, rarely larger groups of records. Their existence was linked with a specific information retrieval need and practice, bound in scope to specific records. On the contrary, our aim is to interconnect different parts of an archive, which were not necessarily meant to be viewed as an ensemble at the time of their creation.

Historical indexes were not only specific in scope, thus adapted to different needs, but were also subject to historical change. Similarly to what Wellisch (1994) pointed out for the indexes of incunabula—that great variety in quality and care involved into index compilation is to be expected—archival indexes are all but uniform. The quality, care and depth of indexation varies tremendously across time and even groups of records. Besides, need stimulates ingenuity, and, as the indexes of printed books improved during the XVI and XVII centuries,

[7]http://iiif.io
[8]https://www.w3.org/TR/annotation-model

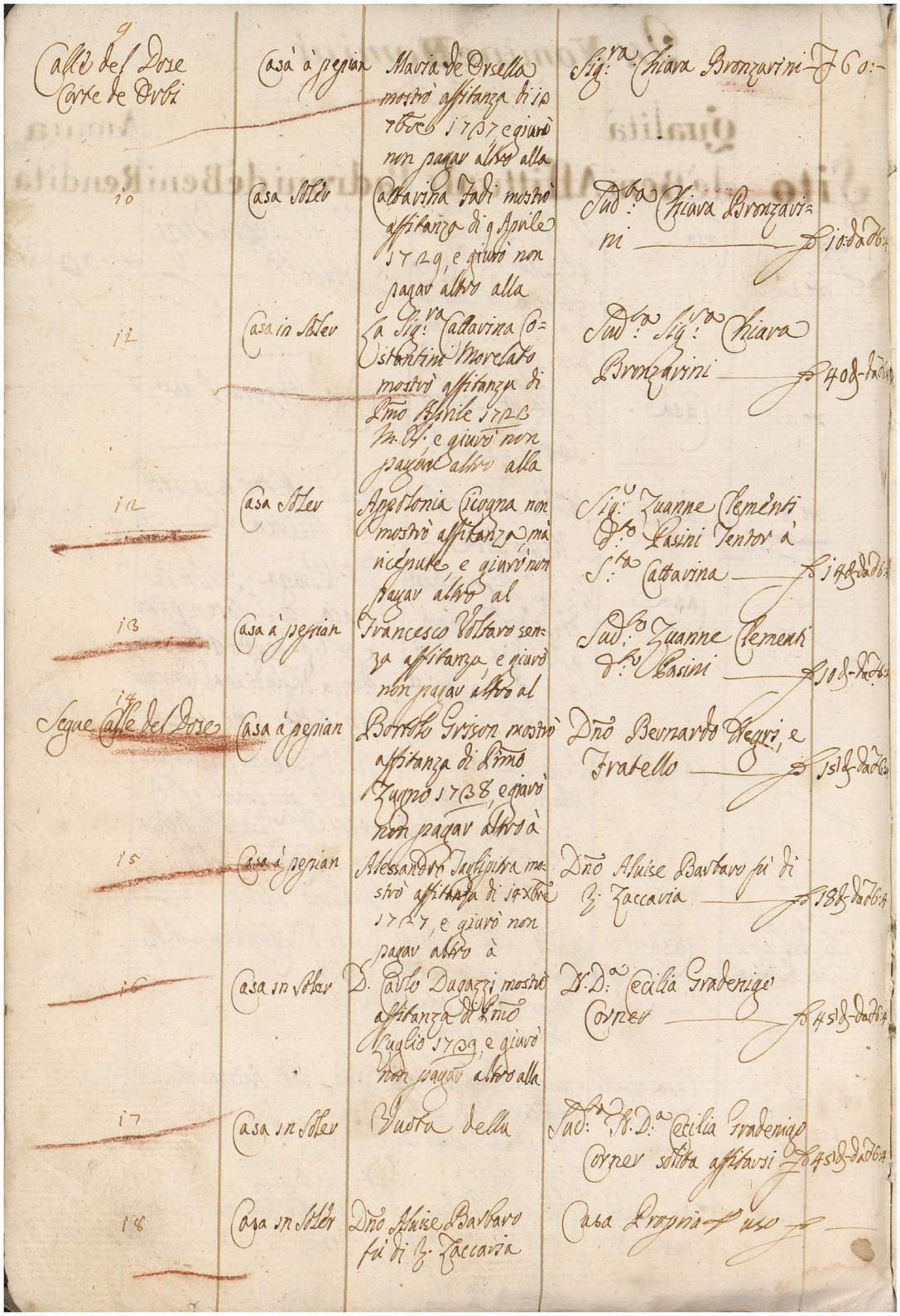

FIGURE 6 | A set of entries in a register of the *Catastici*. Entry 15 (leftmost column, in red) refers to one of the houses owned by "Alvise Barbaro," who is mentioned as landlord in column 4: "Domino Alvise Barbaro fu di Ser Zaccaria." From: State Archive of Venice, *X Savi alle decime in Rialto*, *Catastici delle parrocchie*, 1740, *San Maurizio*, f. 434, c. 891v.

TABLE 1 | Summary of the comparison of the *Quaderni* and *Catastici*.

	Quaderni	***Catastici***
Things indexed	Landlords (juridical person)	Landlords (juridical person)
Coverage (individual entries)	32,406	12,607
Coverage (wrt landlords)	1	1
Coverage (wrt total population)	18.4%	8.7%
Coverage (time span)	1740–1808	1740
Coverage (time)	1	1
Quality (precision)	0.995	0.97
Quality (recall)	0.995	0.96
Overlap	100%	

so did the indexation and management of archival records. It has been recognized that the Western world went through an information overload during the early modern period, sometimes termed "information explosion" (Rosenberg, 2003; Blair, 2011). The phenomenon, largely considered and studied in the context of printed books, might indeed be applicable to archives as well, as shown by early studies on the history of archives, by now a growing field of investigation for historians (De Vivo, 2013; Yale, 2015). It has been suggested that a growth in of the number of records is to be linked with the birth of the first "great archival repositories" of the early modern period, as a consequence of the rise of the "administrative monarchy" (Duchein, 1992, p. 16), and republics perhaps even more (De Vivo, 2013). Furthermore, indexation and other practices designed to cope with information overload in the context of books and learned knowledge might have influenced archive indexation practices, and vice-versa, in a two-way dialog we know was rich (Yale, 2016). The availability of indexes, inventories and other retrieval tools rose with the rise of documentation and the growing interest of states to store and access it during early modern times and beyond (Burke, 2000). Examples abound: in England, the effort of ordering and indexing was often pushed forth by individuals (Popper, 2010). The growing efforts to organize the archive, and make it a pillar of the state, also come from the Swiss Confederation (Head, 2003) and France (Soll, 2009), to name a few. The clear connection between the early modern rising production of records and the introduction of indexation and other record management expedients to cope with it should guarantee the availability of historical indexes for most archives, albeit surveys are lacking at the present time.

Several ways to cope with the task of indexing digitized archives in a more systematic and open manner have been put forward. Crowd-sourcing the historians' community could be an option, and demonstrated by several digital humanities projects (Terras, 2015a). According to Evans (2007), archivists should go through a "shift of values" and accept that item-level descriptions, or means to access the contents of documents and search them, are beyond their reach. They should focus on the structure of the archive, so to say the forest, and leave the contents, or the leaves, to users. Evans' model entails the publication of all records with minimal, collection-level metadata, and the crowd-sourcing of item-level descriptions. The same logic of extensible metadata, that is to say start with metadata at the aggregate level which is then ever-increasing in refinement, is compatible with the index-based method we propose: historical indexes can provide a first set of item-level access, to be complemented in time by other means, for example crowd-sourcing. Another option is the reuse of original metadata, such is the case with historical indexes. A third way to overcome the indexation bottleneck in the "era of digital abundance" is the use of automated methods to extract metadata from records (Yeo, 2013), starting from "barebone" metadata as suggested by Evans (2007). It is likely that a combination of these methods would yield the best results in a setting with increasing digitized records being made available, whilst none in isolation is likely to be sufficient (Yeo, 2013, p. 24). It is worth highlighting how the use of historical indexes, as a way to reuse existing metadata, nicely integrates with a starting point made of a barebone aggregate-level description, and would give an ideal focus to the efforts of the community through crowd-sourcing and automation.

There is hardly any doubt that humanists, and historians among them, make an increasing use of digital tools and resources (Chassanoff, 2013). Even if the number of advanced users among historians is still limited, hardly none is left untouched by the digital turn (Townsend, 2010; Hitchcock, 2013). Nevertheless, digitization rises a set of questions and challenges. Ogilvie (2016) identifies four of them: (i) the decision on what to digitize; (ii) balancing ease of access with privacy and copyright; (iii) what is lost when the archive goes digital, especially concerning its materiality; (iv) what are the possibilities and consequences of being able to create digital collections of documents from multiple archives, instead of accessing them with a traditional, institutional approach. We could add to this list: (v) how are the means of access changing in a digital setting? Questions i, iv, and v are immediately of interest to us. As Ogilvie (2016) readily recognizes, the digitization of historical archives rarely scales to more than a selection of few documents. Nothing comparable to Google Books or the large-scale digitization of historical newspapers by libraries all over the world is yet to be found for archives.

The broader impact of digitization on historians has been the object of some discussion. In a recent article, Putnam (2016) critically explores the new availability of primary and secondary sources online, at a click away through a full-text search, and how this is impacting the way historians work and think. This process has also been called the "Googleization," or "deracination of knowledge" from its traditional holistic environments (Hitchcock, 2013, p. 14). Following Putnam (2016), digital sources offer a disintermediated discovery and foster a transnational research perspective, while previous historians usually acted locally, where their sources were. Yet, digitized sources also come at the risk of loosing contextual awareness that close, slow reading (or, perhaps provocatively, "deep learning") allows to develop. Lastly, there is an increased risk of greatly skewing research in favor of some sources, e.g., printed ones, as has been already shown in the case of digitized newspapers (Milligan, 2013).

6. CONCLUSIONS

The promise of digital and digitized historical archives lays in their indexation at the level of contents. Yet, such indexation is still too slow and costly to scale with the pace of digitization. In this article we presented an approach to prioritize the digitization and bootstrap the indexation of historical archives, using historical indexes as guidance and source of data. We argue that a good way to build the information system backbone of digitized historical archives is to focus on pre-existing indexation tools and align them into unique meta-indexes. The information system thus built will serve as an entry point for the whole archive, as records can be easily connected to the index. Indexes, and consequently groups of records with better or richer indexes, are prioritized during digitization. The proposed approach therefore uses entity-based indexation for the purpose of referential information retrieval. To some extent, the proposal operationalizes what archivists and historians have been doing since a long time: using the context of the archive, and especially its original finding tools, for research. The main steps of the procedure are the archival survey and contextualization, the design of integration, the acquisition and alignment of indexes. We exemplified the proposed method with a case-study: the *X Savi alle Decime in Rialto*, a Venetian magistracy in charge for fiscal administration in early modern Venice. Two indexes giving access to the records of the *X Savi* for the XVIII century were shown to be highly compatible in order to be aligned in a unique meta-index.

The proposed approach has a set of obvious limitations: (i) it is only applicable in the presence of reliable historical indexes; (ii) it entails a digitization and indexation bias toward record groups or document series with historical indexes; (iii) it is still slow, if compared to digitization speed. Limitation (i) is evidently insurmountable, albeit the presence of indexes should be widespread across early modern and modern European archives, as noted previously. Limitation (ii) must be considered in view of the aim of the proposed approach. Indeed, digitization of indexes shall be prioritized but, we argue, then any record which was explicitly or could be practically indexed via the collected index data can be considered for digitization. The method then offers a less biased approach than alternatives. Limitation (iii), albeit true, should be considered in view of the fact that index data is usually more amenable to automatic processing than average. Indexes are normally organized with entries disposed in rows and columns, have few different hands and a relatively stable structure and language, therefore being ideal candidates for automated processing.

We conclude with a more general remark and suggestion for future work. The proposed approach offers one possible way to prioritize the digitization of archival records, namely the presence and quality of existing historical indexes. As noted previously, different quantifiable criteria have been used for the same purpose: for example cost in terms of price or time and popularity among users. We argue that our approach entails a different perspective in that it considers the information contained into records—that is to say their *information potential*—in the specific sense of focusing on high-quality index information. A direction for future work is therefore the exploration of other relevant ways to define and measure the information potential of records, both individually and as a whole, in order to better inform their digitization and indexation.

AUTHOR CONTRIBUTIONS

GC and ME developed the idea, approach and case study, and wrote the paper. FB initially contributed on the idea.

FUNDING

The authors gratefully acknowledge the support of the Lombard-Odier Foundation and the EPFL for the Venice Time Machine project.

ACKNOWLEDGMENTS

This article benefits from the ongoing contribution of the Venice Time Machine team (in alphabetical order): Martina Babetto, Davide Drago, Andrea Erboso, Silvia Ferronato, and Francesca Zugno. The authors would also like to thank the State Archive of Venice and its archivists, especially Paola Benussi and Monica Del Rio, for always insightful guidance. Finally, the authors thank the reviewers and editor for their helpful comments.

REFERENCES

Bailey, J. (2013). Disrespect des fonds: rethinking arrangement and description in born-digital archives. *Arch. J.* Availbe online at: https://web.esrc.unimelb.edu.au/ICAD/bib/P00000019.htm

Beltrami, D. (1954). *Storia della popolazione di Venezia, Dalla Fine del Secolo XVI alla caduta della Repubblica*. Padua: CEDAM.

Blair, A. M. (2011). *Too Much to Know: Managing Scholarly Information Before the Modern Age*. New Haven, CT: Yale University Press.

Boonstra, O., Breure, L., and Doorn, P. (2004). Past, Present and Future of Historical Information Science. *Hist. Soc. Res.* 29, 4–132. doi: 10.12759/hsr.29.2004.2.4-132

Borowiecki, K. J., and Navarrete, T. (2017). Digitization of heritage collections as indicator of innovation. *Econ. Innov. N. Technol.* 26, 227–246. doi: 10.1080/10438599.2016.1164488

Boschetti, F., Cimino, A., DellOrletta, F., Lebani, G. E., Picchi, P., Venturi, G., et al. (2014). "Computational analysis of historical documents: an application to italian war bulletins in world war I and II," in *Proceedings of Workshop on Language Resources and Technologies for Processing and Linking Historical Documents and Archives- Deploying Linked Open Data in Cultural Heritage - LREC* (Reykjavik).

Brin, S., and Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Comput. Netw.* 56, 3825–3833. doi: 10.1016/j.comnet.2012.10.007

Burke, P. (2000). *A Social History of Knowledge: From Gutenberg to Diderot*. Polity Press.

Canal, B. (1908). Il collegio, l'ufficio e l'archivio dei Dieci savi alle decime in Rialto. *Nuovo Archivio Veneto* XVI, 115–150.

Chassanoff, A. (2013). Historians and the use of primary source materials in the digital age. *Am. Arch.* 76, 458–480. doi: 10.17723/aarc.76.2.lh76217m2m376n28

Colavizza, G., Ehrmann, M., and Rochat, Y. (2016). "A method for record linkage with sparse historical data," in *Proceedings of the Digital Humanities Conference* (Cracow).

Coll Ardanuy, M., Knauth, J., Beliankou, A., van den Bos, M., and Sporleder, C. (2016). "Person-centric mining of historical newspaper collections," in *Research and Advanced Technology for Digital Libraries*, Vol. 9819, eds N. Fuhr, L. Kovcs, T. Risse, and W. Nejdl (Cham: Springer International Publishing), 320–331.

De Vivo, F. (2010). Ordering the archive in early modern Venice. *Arch. Sci.* 10, 231–248. doi: 10.1007/s10502-010-9122-1

De Vivo, F. (2013). Coeur de l'Etat, lieu de tension: Le tournant archivistique vu de Venise (XV-XVII siècle). *Ann. HHS* 3, 699–728. Available online at: https://www.cairn.info/revue-annales-2013-3-page-699.htm?contenu=article

Deegan, M., and Tanner, S. (2002). *Digital Futures: Strategies for the Information Age*. Digital Futures Series. Facet Publishing.

DeRidder, J. L., and Matheny, K. G. (2014). *What Do Researchers Need? Feedback on Use of Online Primary Source Materials*. D-Lib Magazine. Available online at: http://www.dlib.org/dlib/july14/deridder/07deridder.html

Duchein, M. (1992). The history of the European archives and the development of the archival profession in Europe. *Am. Archiv.* 55, 14–25. doi: 10.17723/aarc.55.1.k17n44g856577888

Evans, M. J. (2007). Archives of the People, by the People, for the People. *Am. Archiv.* 70, 387–400. doi: 10.17723/aarc.70.2.d157t6667g54536g

Evens, T., and Hauttekeete, L. (2011). Challenges of digital preservation for cultural heritage institutions. *J. Librarianship Inform. Sci.* 43, 157–165. doi: 10.1177/0961000611410585

Grishman, R. (1997). *Information Extraction: Techniques and Challenges*. Berlin; Heidelberg: Springer.

Head, R. (2003). Knowing like a state: the transformation of political knowledge in Swiss archives, 1450-1770. *J. Mod. Hist.* 75, 745–782. doi: 10.1086/383353

Hitchcock, T. (2013). Confronting the digital: or how academic history writing lost the plot. *Cult. Soc. Hist.* 10, 9–23. doi: 10.2752/147800413X13515292098070

Lopatin, L. (2006). Library digitization projects, issues and guidelines: a survey of the literature. *Library Hi Tech* 24, 273–289. doi: 10.1108/07378830610669637

Lynch, C. (2002). Digital collections, digital libraries and the digitization of cultural heritage information. *Microform. Imaging Rev.* 31, 131–145. doi: 10.1515/MFIR.2002.131

Manning, C. D., Raghavan, P., Schütze, H., et al. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press .

Menini, S., Sprugnoli, R., Moretti, G., Bignotti, E., Tonelli, S., and Lepri, B. (2017). "Ramble on: tracing movements of popular historical figures," in *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Valencia: Association for Computational Linguistics), 77–80.

Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., et al. (2015). Semantic technologies for historical research: a survey. *Semant. Web* 6, 539–564. doi: 10.3233/SW-140158

Milligan, I. (2013). Illusionary order: online databases, optical character recognition, and Canadian history, 1997-2010. *Can. Hist. Rev.* 94, 540–569. doi: 10.3138/chr.694

Moss, M., Thomas, D., and Gollins, T. (2018). Artificial fibers - the implications of the digital for archival access. *Front. Dig. Hum.* 5:20. doi: 10.3389/fdigh.2018.00020

Nauta, G. J., and van den Heuvel, W. (2015). *Survey Report on Digitisation in European Cultural Heritage Institutions.* Technical Report, Europeana/ENUMERATE.

Ogilvie, B. (2016). Scientific archives in the age of digitization. *Isis* 107, 77–85. doi: 10.1086/686075

Olieman, A., Beelen, K., van Lange, M., Kamps, J., and Marx, M. (2017). "Good applications for crummy entity Linkers? The case of corpus selection in digital humanities," in *Proceedings of the 13th International Conference on Semantic Systems* (Amsterdam: ACM).

Ooghe, B., Waasland, H. C., and Moreels, D. (2009). Analysing selection for digitisation. *D-Lib Magazine* 15, 1082–9873. doi: 10.1045/september2009-ooghe

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. Morgan and Claypool Publishers.

Piskorski, J., and Yangarber, R. (2013). *Information Extraction: Past, Present and Future*. Berlin;Heidelberg: Springer.

Popper, N. (2010). From abbey to archive: managing texts and records in early modern England. *Arch. Sci.* 10, 249–266. doi: 10.1007/s10502-010-9128-8

Putnam, L. (2016). The transnational and the text-searchable: digitized sources and the shadows they cast. *Am. Hist. Rev.* 121, 377–402. doi: 10.1093/ahr/121.2.377

Rikowski, R. (2008). Digital libraries and digitisation: an overview and critique. *Policy Fut. Educ.* 6, 5–21. doi: 10.2304/pfie.2008.6.1.5

Rosenberg, D. (2003). Early modern information overload. *J. Hist. Ideas* 64, 1–9. doi: 10.1353/jhi.2003.0017

Rovera, M., Nanni, F., Ponzetto, S., and Goy, A. (2017). "Domain-specific named entity disambiguation in historical memoirs," in *Proceedings of the Fourth Italian Conference on Computational Linguistics 742 CLiC-it 2017*, eds R. Basili, M. Nissim, and G. Satta (Rome: Accademia University Press), 287–291.

Soll, J. (2009). *The Information Master: Jean-Baptiste Colbert's Secret State Intelligence System*. The University of Michigan Press.

Tanner, S. (2006). *Handbook on Cost Reduction in Digitisation*. Minerva Project.

Terras, M. (2011). "The rise of digitization," in *Digitisation Perspectives*, Vol. 39, ed R. Rikowski (Rotterdam: SensePublishers), 3–20.

Terras, M. (2015a). "Crowdsourcing in the digital humanities," in *A New Companion to Digital Humanities*, eds S. Schreibman, R. Siemens, and J. Unsworth (Chichester: John Wiley & Sons, Ltd.), 420–438.

Terras, M. (2015b). Opening access to collections: the making and using of open digitised cultural content. *Online Inform. Rev.* 39, 733–752. doi: 10.1108/OIR-06-2015-0193

Theimer, K. (2012). Archives in context and as context. *J. Digit. Human.* 1, 1–8. Available online at: http://journalofdigitalhumanities.org/1-2/archives-in-context-and-as-context-by-kate-theimer/

Townsend, R. B. (2010). How is the New Media Reshaping the Work of Historians? Available online at: https://www.historians.org/publications-and-directories/perspectives-on-history/november-2010/how-is-new-media-reshaping-the-work-of-historians

Wellisch, H. H. (1994). Incunabula indexes. *Indexer* 19, 3–12.

Yale, E. (2015). The history of archives: the state of the discipline. *Book Hist.* 18, 332–359. doi: 10.1353/bh.2015.0007

Yale, E. (2016). The book and the archive in the history of science. *Isis* 107, 106–115. doi: 10.1086/686078

Yeo, G. (2012). The conceptual fonds and the physical collection. *Archivaria* 73, 43–80. Available online at: https://archivaria.ca/archivar/index.php/archivaria/article/view/13384

Yeo, G. (2013). Archival description in the era of digital abundance. *Comma* 2, 15–26. doi: 10.3828/comma.2013.2.2

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.