The Impact of Digital Editing on the Study of Holocaust Survivors' Testimonies in the context of Voci dall'Inferno Project

Angelo Mario Del Grosso, Marina Riccucci, Elvira Mercatanti

CNR-ILC, UNIPI,

Via Moruzzi, Pisa; Piazza Dante, Pisa

angelomario.delgrosso@cnr.it, marina.riccucci@unipi.it, e.mercatanti@studenti.unipi.it

Abstract

In Nazi concentration camps, approximately 20 million people perished. This included young and old, men and women, Jews, dissidents, and homosexuals. Only 10% of those deported survived. This paper introduces "Voci dall'Inferno" project, which aims to achieve two key objectives: a) Create a comprehensive digital archive: by encoding a corpus of non-literary testimonies including both written and oral sources. b) Analyze the use of Dante's language: by identifying the presence of Dante's lexicon and allusions. Currently, the project holds 47 testimonies, with 29 transcribed in full text and 18 encoded using the XML-TEI format. This project is propelled by a multidisciplinary and educational context with experts in humanities and computer science. The project's findings will be disseminated through a user-friendly web application built on an XML foundation. Though currently in its prototyping phase, the application boasts several features, including a search engine for testimonies, terms, or phrases within the corpus. Additionally, a browsing interface allows users to read and listen the original testimonies, while a visualization tool enables deeper exploration of the corpus's content. Adhering to the Text Encoding Initiative (TEI) guidelines, the project ensures a structured digital archive, aligned with the FAIR principles for data accessibility and reusability.

Keywords: XML-TEI, Holocaust Testimonies, Digital Archives

1. Introduction

The concentration camps we know are the camps of 1944 because almost none of the *Häftlinge*, namely the prisoners, survived to tell us about the other camps[1]. In the last years, many of the survivors have died. In addition, not all testimonies have told their story, but only a part, certainly not the majority. Many survivors have remained silent. Moreover, the *Camp* is *ineffable* (Levi, 2018, p. 688). Thus, for the testimonies, the problem was also to find the words to express the atrocity they suffered (Wiesel, 1995). To tell the story of the Camp, therefore, one must overcome the poverty of a vocabulary that does not have "*the words to say it*".

Saying that the *Lager* was *l'Inferno - the hell*[2] (Calderini and Riccucci, 2020) has allowed survivors to establish an immediate contact with their audience. This shared metaphor always recurs in holocaust testimonies.

Since 2016, professorship of Italian Literature at the University of Pisa has directed and coordinated, with the support of the CNR-ILC of Pisa, the research project called "Voci dall'Inferno". This scholarly initiative has two integrated and correlated objectives:

a) The digitization and encoding of a corpus, the largest possible, of testimonies.

b) The identification, quantification and evaluation of the presence of Dante's lexicon and imagery within those testimonies.

Early outcomes have been surprisingly important: given the vast scope of the project, the very high number of testimonies, most of which are unpublished, it has been necessary to involve computational methods and techniques[3] as well as to make the research activities a collaborative work, which has seen and continues to see the collaboration and contribution of many DH students.

2. Testimonies and Dante's Lexicon

The textual typologies through which the Lager has been reported to us *in words* are essentially two:

- Direct testimony - coeval and non-coeval - of those who experienced the extermination

---

[1] Encyclopedia of the Holocaust: https://www.ushmm.org/it

[2] It is worth noting that the Inferno here refers to Dante's literary work *Divina Commedia* (Alighieri, 2002), not the Christian concept of Hell

[3] Currently, the corpus is manually annotated and digital encoded. This initial, time-consuming effort will serve as a valuable ground truth for training and refining sophisticated machine learning and prediction tools for automated data extraction and encoding. Examples include leveraging current AI applications for tasks such as handwritten text recognition, automatic speech recognition, named entity recognition, topic extraction and modeling, semantic text alignment, and more. Once the final topics are defined, drawing upon the work done on similar collections, we also plan to incorporate formal ontologies and linked open data.

camp and reported it in forms that only rarely touch on literariness: the *modus dicendi* of this type of testimony is located in the space between the oral report (the interview) and the written one (the diary, the autobiographical/memorial story, the letter).

- Indirect testimony - coeval and non-coeval - of those who experienced the extermination camp and chose, to report it the form of narrative, therefore of literature. This kind of testimony is presented in the form of an organized story, thematically and stylistically structured.

The latter category encompasses works with manifested literary ambitions that would never have been written if the Holocaust and the deportation had not occurred. We are referring to that literary production like *Se questo è un uomo - If this is a man* (Levi, 1947), *Night* (Wiesel, 1960), *This Way for the Gas, Ladies and Gentlemen* (Borowski, 1976), *The Human Race* (Antelme et al., 1998), *L'Univers concentrationnaire* (Rousset, 1946), *Fatelessness* (Kertész, 1975). In the context of our research project we call this kind of testimonies "second level" ones.

We are faced with two distinct and differentiated forms of representation/restitution of the concentrationary universe. What has been progressively verified is not only that the lexicon of the Inferno of Dante breaks the silence and intervenes in those who are going to narrate the tragedy of their experience (Arquès, 2009; Pertile, 2010; Susteric, 2016; Taterka, 2002) providing the words "to say it", but also that the expressive faculties of all witnesses are influenced by the Dante's Inferno, even of those who have read Dante only on the school benches, or, even, of those who have received Dante as a heritage of oral culture.

By tacit agreement, all the *Häftlinge*, without exception, who have testified have chosen and adopted the term "*Inferno*" to convey their experience of the concentration camp to those who have never experienced it themselves.

There are at least two basic assumptions from which we cannot prescind:

1. The metaphorical nexus (Lager-Inferno) is a new nexus. It did not exist before, simply because before the 30s of the twentieth century, the Lager did not exist.

2. When the testimonies talk about the Lager as of the Inferno they do not refer to any inferno, or to an inferno and nothing else. They do so with the Dante's inferno in mind.

The data collected so far tell us that Dante breaks the silence, in the sense that survivors use Dante's lexicon to untie the knot of ineffability, to dilute the paralysis of mind and memory in the face of the emergence of the nefarious experience. This happens in all survivors (Riccucci and Riccotti, 2021).

Listening to and reading the testimonies of the Nazi camps, one realizes that to report on the concentrationary hell, the survivors of the extermination, people of every level of education, relied on the words of the Dante's vocabulary, that of the first canticle, for the most part, that of the Inferno. Of course, the testimonies are not all and not systematically or capillarily woven with Dante's verses, but these verses at a certain point burst from the lips of these men and women to express the inexpressible.

The testimonies can count on a lexical heritage and a collective imagery made up of Dante's words that have entered into common use, penetrated into the language of everyday life, into the speech of the family and of society, transmitted from generation to generation as an inheritance.

In the light of the above, computer science allows us to preserve and archive, to interrogate and find connections, to build maps, to intertwine stories[4].

3. Digital Archive and Corpus

The Voci dall'Inferno project has undergone three distinct phases, each contributing to its current state as a web platform for studying Holocaust testimonies[5]:

1. Development of a database for the management of the testimony records - the archive has been named *memoria-archivio* (Riccucci et al., 2021).

2. Creation of the corpus of the testimonies in XML-TEI format (Burnard, 2014).

3. Development of a web application for the presentation and interrogation of the data stored in the digital archive.

---

[4]Similar initiatives encompass projects like *Let Them Speak*, https://lts.fortunoff.library.yale.edu/, or *David Boder: from wire recordings*, https://ranke2.uni.lu/u/boder/, or archives like *CDEC*, https://digital-library.cdec.it/cdec-web/, as well as initiatives like *EHRI*, https://www.ehri-project.eu/. Other projects of a similar nature can be found at https://dhjewish.org/projects.

[5]The corpus - now encompassing about 500,000 tokens in Italian - is not yet exhaustive of all available primary sources. We plan to incorporate new oral and written testimonies during our research. The initial corpus is conceived as the first iteration of the project, designed to primarily establish the data model, the encoding schema, the workflow, and the scholarly framework. During this initial phase of data collection, resources were gathered from private and personal archives, public institutional archives, audio type interviews, and video interviews

The *memoria-archivio* database provides us with a web environment for the creation of the initial inventory of the testimonies. It preserves catalogographic and literary descriptions as well as, where present, also manages the transcriptions of the textual content of the testimony. In addition, the web environment allows for the comparison of the testimony lexicon with the text of the Dante's Comedy.

The application allows for the updating of the inventory, the records of the witnesses and the curators of the sources. Subsequently, memoria-archivio also integrated the management of documents in XML-TEI format. To date, the archive has 47 testimonies, of which 29 are full-text transcriptions and 18 documents has been also encoded in XML-TEI format (Fig. 1). These collection of testimonies includes known and lesser-known names of Italians who were deported, such as Samuele Modiano, Piero Terracina, Enrico Vanzini, Liliana Segre, Nedo Fiano, Shlomo Venezia, Primo Levi, Ida Marcheria, Goti Bauer, and many others.

Within the digital archive, the testimonies are divided into two macro-classes, which determine their representational, functional and exploitation aspects. On the one hand, there are the oral testimonies (Fig. 6) and on the other hand, the written testimonies (Fig. 2). While maintaining the specific differences, both classes follow the guidelines provided by the TEI consortium.

In particular, during the course of the project, a "One Document Does it All" (ODD)[6] was created - and gradually refined, which declares the modules, elements, attributes and possible values allowed for the encoding of the corpus. As for the encoding model, the written testimonies follow an image-based editing scheme of a diplomatic-interpretative type with a parallel-transcription approach to the representation of the text-document (Pierazzo, 2015). These kinds of digital documents can be published by means of the EVT Web application (Rosselli Del Turco et al., 2019).

To this end, the elements defined in module 11 (transcription) were used for the transcription of the primary source; module 13 (namesdates) for the representation of named entities; module 16 (linking) for the analysis of particular text structures; module 17 (analysis) for the semantic and linguistic-lexical annotation of textual units. As for the description of the primary source, the elements defined in module 10 (msDesc) of the TEI guidelines were adopted.

3.1. Metadating the Testimonies

Oral testimonies differ structurally from written ones due to the inherent characteristics of spoken language. Notably, oral testimonies are characterized by the temporal dimension of speech and the unique order in which utterances unfold.

This leads to the creation of specific "timeline" elements designed to synchronize the topics covered by the witness. This synchronization aims to align the timing of the speaker's utterances with their corresponding transcription.

```
<abstract><ab><list>
 <item synch="#TR1">
  <persName ref="#LS">
   <forename>Liliana</forename>
   <surname>Segre</surname>
  </persName>, rispondendo alla domanda
   postale da <persName ref="#AS">
   <forename>Anna</forename>
   <surname>Segre</surname>
  </persName>, parla di che cosa abbia
 </item>
</abstract>

<standOff>
 <timeline xml:id="TL1I"
 source="#reg_1B" unit="s">
 <!-- ... -->
 <when xml:id="TR1"
  absolute="00:00:00"/>
 <!-- ... -->
 <when xml:id="TR7"
  absolute="00:23:41"/>
 <!-- ... -->
 </timeline><!-- ... -->
</standOff>
```

In addition to the timeline, a section of reasoned summary, called *regesto*, has been introduced, which briefly illustrates the content of each division. Four different timelines have been defined in the encoding model.

In the first timeline, the ‹when/› elements identify the moments when the various topics of the testimony are introduced, i.e. those summarized within the ‹item› elements present in the *regesto*, which, in turn, have a @synch attribute in order to connect the timing specified by the relative ‹when/› tag.

In the second timeline, on the other hand, all the segments in which the voices overlap have been collected. For this reason, the ‹when/› elements inside it are in pairs: one of them identifies the interval of the overlap.

The third timeline was instead created to group together the moments in which a change of speaker occurs. The fourth and last timeline allows to record the moments in which background noises overlap with the utterances.

The most significant XML-TEI elements used for the description of oral sources can be summarized as follows: the information relating to the medium

---

Figure 1: Memoria-Archivio Web Application

Figure 2: EVT Web Application for Autographs' Primary Sources

was recorded using the <recordingStmt> element, contained in turn in the <sourceDesc> element, belonging to module 8 of the TEI guidelines (spoken). The <recording> element finally represents a single recording and contains all the information necessary to specify the context and responsibilities of the recording. Each <recording> element is connoted with the type attribute (@type) (audio or video) and a duration (@dur) for each single recording.

```
<sourceDesc>
 <recordingStmt>
```

```
<recording type="audio"
 dur="P30M41S" xml:id="reg_1B">
 <p>Registrazione 1 lato B</p>
 <date cert="low"
   when="2007-12-08">
   8 dicembre 2007</date>
</recording>
<recording type="audio"
 dur="P30M41S" xml:id="reg_2A">
 <p>Registrazione 2 lato A</p>
 <date cert="low"
   when="2007-12-08">
```

```
    8 dicembre 2007</date>
  </recording>
  <recording type="audio"
   dur="P24M32S" xml:id="reg_2B">
   <p>Registrazione 2 lato B</p>
   <date cert="low"
     when="2007-12-08">
     8 dicembre 2007</date>
  </recording>
 </recordingStmt>
</sourceDesc>
```

Finally, within the ⟨profileDesc⟩ block, the elements defined in module 15 (corpus) were used, in particular the ⟨particDesc⟩ element offers an accurate description of the people who took part in the conversation.

3.2. Transcribing the Testimonies

The two encoding models introduced in previous sections, namely the model for oral testimonies and the model for written testimonies, differ from each other both in terms of descriptive and analytical choices.

The logical structure of the written testimony often follows a predominantly epistolary grammar, but it can differ substantially due to authorial and editorial characteristics (authorial interventions on manuscripts or typewritten texts).

The primary source is represented using the TEI facsimile tagset to describe areas of interest through relevant attributes within the @zone element. These zones are then referenced by corresponding elements within the transcription section, adhering to the best practices outlined by the parallel-transcription method (TEI module 11)[7].

Within the transcription section (⟨text⟩ block), various elements are used to represent the content of the primary source. These include:

- ⟨subst⟩ or ⟨mod⟩: These elements mark the original text with authorial interventions.

- ⟨choice⟩ elements with their sub-elements: These elements record the original reading and any editorial interpretations made by scholars.

- ⟨damage⟩, ⟨unclear⟩ and ⟨supplied⟩: This element allows scholars to include missing or unclear information, providing any difficulties encountered due to material damage.

Named entities encoding follows the best practices suggested by the TEI guidelines, adopting the elements ⟨person⟩, ⟨org⟩, ⟨place⟩, ⟨event⟩ and the respective ⟨personName⟩, ⟨orgName⟩, ⟨placeName⟩. Dante's quotes and terminology have been annotated with the ⟨cit⟩ and ⟨term⟩ elements respectively.

As for oral testimonies, the transcription is divided into textual units called utterances using the ⟨u⟩ element. Each utterance is accompanied by the @who attribute which allows to associate the person who formulated it. In addition, the @xml:id and @synch elements are functional for a correct synchronization. The @trans attribute, on the other hand, specifies whether the participants' utterances follow one another or overlap.

During the testimony, numerous phenomena are recorded, such as pauses (element ⟨pause⟩ with attribute @type to indicate the length), non-lexical sounds (element ⟨vocal⟩), prosodic events (element ⟨kinesic⟩), background noises (element ⟨incident⟩), inaudible or uncertain passages (elements ⟨gap⟩ and ⟨unclear⟩), changes in paralinguistic features such as intonation, volume, rhythm, speed by means of the ⟨shift /⟩ element accompanied, as is appropriate, by the attributes @feature and @new.

```
<u><!-- ... --!>
non hai un nome, perché
hai un numero,
<pause type="long"/>
ti chiamano per numero
<pause type="medium"/>
e quindi <pause type="short"/>
cercano di
degradarti <pause type="short"/>
con la fame <pause type="short"/>
<!-- ... ---!> </u>

<u who="#MARCHERIA" xml:id="m223"
  synch="#tlp457"> In questo
 <supplied>caos</supplied>
 sì, perché arrivavano i russi
 <pause type="short"/>
 e c'era il caos. Siamo
 <del type="repetition">siamo</del>
 <kinesic>
 <desc>
 Ida mostra la grandezza della piazza
 </desc>
</kinesic>
```

4. Voci Dall'Inferno Web Applet

The digital archive of testimonies would be less effective from a functional and scientific point of view without the presence of a software component dedicated to the extraction, manipulation, presentation and use of the data collected during the encoding phase.

Figure 3: Voci dall'Inferno Web Application developed with Saxonjs2

Two different data restitution strategies were experimented, which make use of two different architectural approaches.

For the first approach, web applications were developed by leveraging the functionalities of a client-side library for processing XML documents, namely SaxonJS2[8] (Fig. 4).
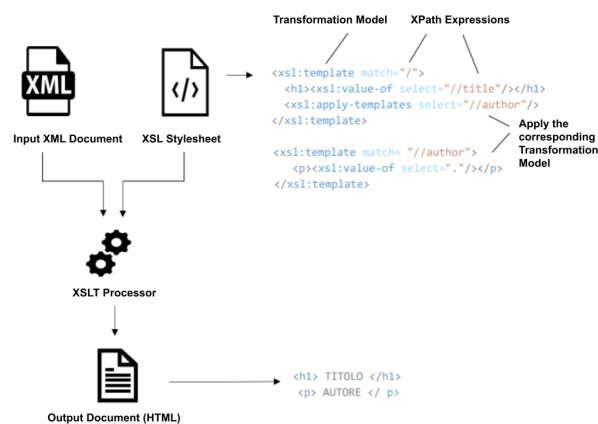

Figure 4: Saxon Processing Model

In relation to the second approach, the web applications were developed by means of the eXist-db[9] environment using the HTML templating module (server-side)[10].

Thanks to the use of the SaxonJS2 library, it is possible to integrate an efficient XSLT processor by delegating the browser's javascript engine to manipulate the DOM object of the HTML page.

The library exposes an effective API whose main methods are `SaxonJS.transform (options)`
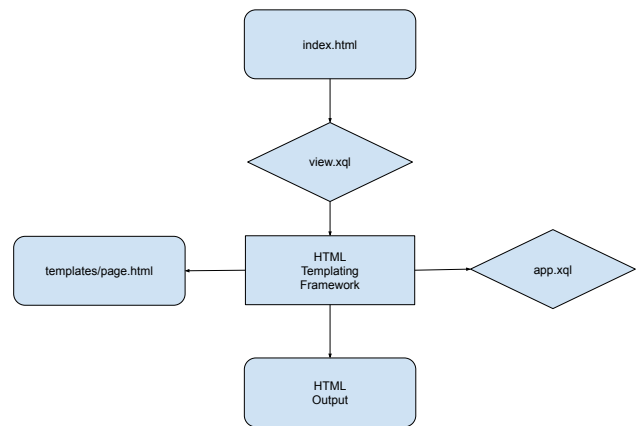
Figure 5: eXist-db Processing Model with HTML Templating

to execute the procedures defined by the transformation rules and `SaxonJS.XPath.evaluate(XPath)` to select appropriate sequences of XML nodes or process them according to the specifications of the XPath 3.1 standard.

```
SaxonJS.getResource({
  location: "testimony.xml",
  type: "xml" })
  .then(doc => {
    const result =
     SaxonJS.XPath
     .evaluate(
      "//persName/text()", doc);
    const output =
     SaxonJS.serialize(result, {
    method: "xml",
    indent: true,
    "omit-xml-declaration":true});})
```

The image in Figure 3 shows a web page generated with the help of the SaxonJS2 library for

---

[8]See also at https://www.saxonica.com/saxon-js/index.xml

[9]See also at https://exist-db.org/exist/apps/homepage/index.html

[10]See also at https://github.com/eXist-db/templating

6

the visualization of the testimony of Arminio Wachs-berger. It is possible to notice the transcribed text of the interview, the participants, the textual phenomena annotated and rendered graphically according to the styles indicated in the legend.

The second approach is based on the eXist-db technology. As introduced, the platform integrates a module dedicated to the dynamic generation of HTML pages starting from collections of documents in XML format and from procedures implemented using the XQuery instructions. The basic operation involves the use of HTML templates, in which appropriate directives and calls to XQuery functions are added. The functions implement the application logic to generate the HTML fragments useful to complete the actual HTML page.

A relevant feature of the eXist-db technology is the possibility to use the Apache Lucene library for indexing textual data and for the efficient interrogation of them. Figure 7 shows an example of querying and retrieving data related to the testimony of Ida Marcheria (partial word search "tren").
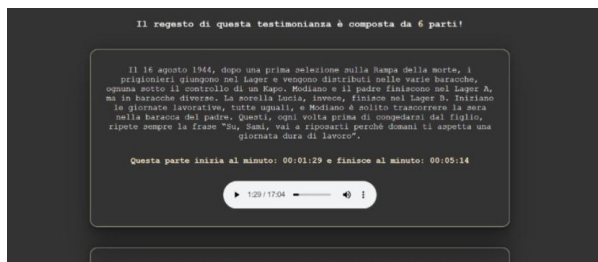
Figure 6: Voci dall'Inferno Web Application for Audio and Regesto Features

The web application developed so far for Voci dall'Inferno (Fig. 8) has multiple functionalities implemented or under development such as: I) Catalog management and search (Fig. 1); II) Presentation and use of data in parallel with the primary source (Fig. 2); III) Search within the textual archive (Fig. 7); IV) Management of the regesto (Fig. 6); V) Management of speech conventions (Konrad, 2003) (Fig. 8); VI) Statistics of phenomena (Fig. 3); VII) Terminology management (Fig. 7); VIII) Management of quotations and allusions (Dante's ones in particular).

5. Conclusion

We have presented the *Voci dall'Inferno* project, a scholarly initiative, within an educational framework, building a digital corpus of Holocaust testimonies and a dedicated digital environment for searching and analyzing them. This project uniquely explores the presence of Dante's vocabulary and allusions within the testimonies. The digital collection adheres to the *Text Encoding Initiative* (TEI) schema,

maximizing data accessibility, searchability, interoperability, and reusability in accordance with the *FAIR principles* (Wilkinson and al., 2016). Notably, the project is developing functionalities for data classification and extraction using machine learning techniques. These functionalities will enable automatic speech recognition and transcription, automatic search for literary tesserae, as well as automatic network analysis, further enriching the exploration and understanding of these testimonies.

6. Bibliographical References

Dante Alighieri. 2002. *La divina commedia. Inferno*. Armando Editore.

Robert Antelme, Jeffrey Haight, and Annie Mahler. 1998. *The Human Race*. Marlboro Press/Northwestern.

Rossend Arquès. 2009. Dante nell'inferno moderno:la letteratura dopo auschwitz. *Rassegna Europea di Letteratura Italiana*, 33.

Tadeusz Borowski. 1976. *This way for the gas, ladies and gentlemen*, 1st ed. edition. Penguin Books, New York.

Tadeusz Borowski. 2009. *Da questa parte, per il gas*. Un mondo a parte. L'Ancora del Mediterraneo.

Lou Burnard. 2014. *What Is the Text Encoding Initiative?: How to Add Intelligent Markup to Digital Resources*. OpenEdition Press, Marseille.

Sara Calderini and Marina Riccucci. 2020. Le parole per dire il lager. *Italianistica*, XLIX.

Imre Kertész. 1975. *Fatelessness*. Vintage.

Ehlich Konrad. 2003. Hiat: A transcription system for discourse data. *Talking Data: Transcription and Coding in Discourse Research*.

Primo Levi. 1947. *Se questo è un uomo*. De Silva, Torino.

Primo Levi. 2018. *Opere Complete*. Giulio Einaudi editore.

Lino Pertile. 2010. L'inferno, il lager, la poesia. *Dante*, 7.

Elena Pierazzo. 2015. *Digital Scholarly Editing: Theories, Models and Methods*. Ashgate, Surrey.

Marina Riccucci, Angelo Mario Del Grosso, Frida Valecchi, and Giulia Causarano. 2021. Testimoniare il lager: L'informatica al servizio della memoria. *Quaderni Di Umanistica Digitale*.

Marina Riccucci and Laura Riccotti. 2021. *Il dovere della parola. Le testimonianze di Liliana Segre e di Goti Herskovits Bauer*, volume XLIX. Pacini editore, Pisa.

Roberto Rosselli Del Turco, Chiara Martignano, Chiara Di Pietro, Giulia Cacioli, Angelo Mario Del Grosso, and Simone Zenzaro. 2019. DSE Visualisation with EVT: Simplicity is Complex. In *Complexities*.

David Rousset. 1946. *L'Univers concentrationnaire*. éditions du Pavois.

Federica Susteric. 2016. La dicibilità del male. la ricezione dantesca nelle testimonianze concentrazionarie. *Dante*, XII.

Thomas Taterka. 2002. Dante deutsch. *Studi sulla letteratura dei Lager*.

Elie Wiesel. 1960. *Night*. Hill 'n' Wang, New York.

Elie Wiesel. 1995. *La notte*. Giuntina editore.

Mark Wilkinson and al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Figure 7: Voci dall'Inferno Web Application - Search Capabilities

Figure 8: Voci dall'Inferno Web Application - Current Implementation in eXist-db