Noname manuscript No.
(will be inserted by the editor)

A linked open data architecture for the historical archives of the Getulio Vargas Foundation

Alexandre Rademaker · Dário Augusto Borges Oliveira · Valeria de Paiva · Suemi Higuchi · Asla Sá · Moacyr Alvim

the date of receipt and acceptance should be inserted later

Abstract This paper presents an architecture for historical archives maintenance based on Open Linked Data technologies and open source distributed development model and tools. The proposed architecture is being implemented for the archives of the Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC, Center for Research and Documentation of Brazilian Contemporary History) of the Fundação Getulio Vargas (FGV, Getulio Vargas Foundation). We discuss the benefits of this initiative and suggest ways of implementing it, as well as describing the preliminary milestones already achieved. We also present some of the possibilities for extending the accessibility and usefulness of the data archives information using semantic web technologies, natural language processing, image analysis tools, and audio-textual alignment, both in progress and planned.

Alexandre Rademaker
IBM Research and FGV/EMAp
alexrad@br.ibm.com

Dário Augusto Borges Oliveira
FGV/CPDOC
darioaugusto@gmail.com

Valeria de Paiva
Nuance Communications
valeria.depaiva@gmail.com

Suemi Higuchi
FGV/CPDOC
Suemi.Higuchi@fgv.br

Asla de Sá
FGV/EMAp
asla.sa@fgv.br

Moacyr Alvim
FGV/EMAp
Moacyr.Silva@fgv.br

Keywords historical archives · digital humanities · Semantic Web · NLP · image processing · audio processing · open data.

1 Introduction

The paradigm of linked open data has changed significantly the way knowledge and information are made available over the Internet. Information portals are facing the challenge of enhancing semantically their information so as to provide richer and interlinked content, which ultimately allows users to access data more efficiently for their specific applications.

Institutions that provide content with recognized quality, such as universities and museums, are specially interested in having their rich data accessed and referenced by a broader audience.

In the field of collections, heritage and cultural assets, we can identify many efforts to publish existing metadata as linked open data. The BiographyNet [19] is extremely similar to our project. It is a multi-disciplinary project bringing together history, Linked Data and tools that aims at enhancing the research potential of the Biography Portal of the Netherlands [1], a heterogenous collection made up out of 23 sources which provides access to over 125,000 entries describing 76,000 people considered prominent figures of Dutch history.

Other similar projects are: the Europeana project [23, 39] which mapped and published data from more than 2.000 institutions across Europe; the Smithsonian project [43] which published data from a collection of 41.000 objects from the Smithsonian American Art Museum; and the Finnish Museums project [24] which published data concerning some 260 historical sites in Finland. These initiatives aim at promoting the integration of digital collections of cultural heritage

[1] http://www.biografischportaal.nl/en

based on the use of archival metadata, cross-domain ontologies and open data technologies [20].

Even though much public data is available freely online in Brazil, only few repositories use open data standards. Examples in this direction are the Federal Government Open Data [17], the LeXML [29] and the SNIIC [13] projects. Despite being a reference in the field of organizing and preserving historical collections, CPDOC (the Center for Research and Documentation of Brazilian Contemporary History of the Getulio Vargas Foundation) currently does not adopt any metadata standards nor does it use any open data model for its collections.

Given the trends for data sharing and interoperability of digital collections, it is a challenge to keep CPDOC innovative in its mission of efficiently making available historical data. In this article, we present CPDOC's collections and discuss our approach for delivering its content using semantics technologies. We also present some audiovisual signal and natural language processing tools that we are using to enrich the metadata of the documents, to allow better search and browsing experience on the collections. Our proposal introduces changes in the way CPDOC deals with its archives maintenance and accessibility, building a model for data organization and storage that ensures easy access, interoperability and reuse by service providers. The goal is to deliver an open and flexible framework, that uses semantically interconnected data about images, audio and textual content, to provide knowledge in a smart, collaborative and efficient environment.

Concerning automatic extraction of semantics from the collections, three major applications are being currently explored and will be described in the next sections: (1) natural language processing for enriching the metadata and extracting knowledge currently embedded into the historical dictionary's textual entries; (2) voice recognition and transcription alignment from the audiovisual archives of oral history interviews; and (3) face detection and identification of important characters in historical photographs [44]. Among other details, this paper complements [40] with the description of these techniques and how we use them to expand the metadata from knowledge extraction from raw data in images, sound and text files.

Amongst our main objectives are the construction of a RDF [32] data store from data originally stored in a relational database and the construction of an OWL [16] ontology to properly represent the CPDOC domain, using Open Linked Data Initiative principles [26]. The project also aims to make the RDF data freely available for downloading, similarly to what DBpedia [6] does. We believe that these efforts will motivate the interoperability of CPDOC's collections with other open data projects.

The paper is organized as follows: Section 2 presents CPDOC's current database and information systems architecture. The shortcomings of this architecture are described in Section 3 and the requirements for a new one in Section 4. The new architecture proposed in this work is presented in Section 5, and in Section 6 we exploit the possibilities of enhancing the data in the archives using complementary technologies. In Section 7 we present some plans for evaluating our proposal and the results obtained. Finally, conclusions are summarised in Section 8.

2 Current database and information systems architecture

CPDOC was created in 1973 and became an important historical research institute in Brazil, housing a major collection of personal archives, oral history interviews and audiovisual sources. Since its foundation, the center has received the donation of personal archives of prominent Brazilian figures from the 1930s onward, starting with President Getulio Vargas himself. In 1975 the institute launched its Oral History Program (PHO), which involved the recording and archiving of interviews with major players in events in Brazilian history. In 1984 the center published the Brazilian Historical-Biographical Dictionary [1] (DHBB), a regularly updated reference resource that documents the contemporary history of the country.

More recently the center has placed an increasing emphasis on applied research, working in collaborative projects that extend the availability and scope of the valuable historical records it holds.

The current CPDOC database architecture is presented in Figure 1. The data is stored in three different information systems that share a common relational database. Each of the systems is maintained separately and adopts idiosyncratic criteria concerning the organization and indexing of its information, which vary depending on the specifications of the content it hosts: personal archives documents, oral history interviews and the Brazilian Historical Biographic Dictionary entries. CPDOC's website provides a query interface to these data. In the following subsections we briefly describe each of the systems.

2.1 Personal Archives (Accessus)

This system contains information from personal files of people who influenced onwards scene in Brazil from the early 20th century. These historical documents, in textual or audiovisual sources, represent more than private memories, they are registries of a collective memory.

Currently, more than 200 personal archives from presidents, ministers, military personal and others constitute the Accessus collections. The organization structure of the collections follows the guidelines established for archiving and

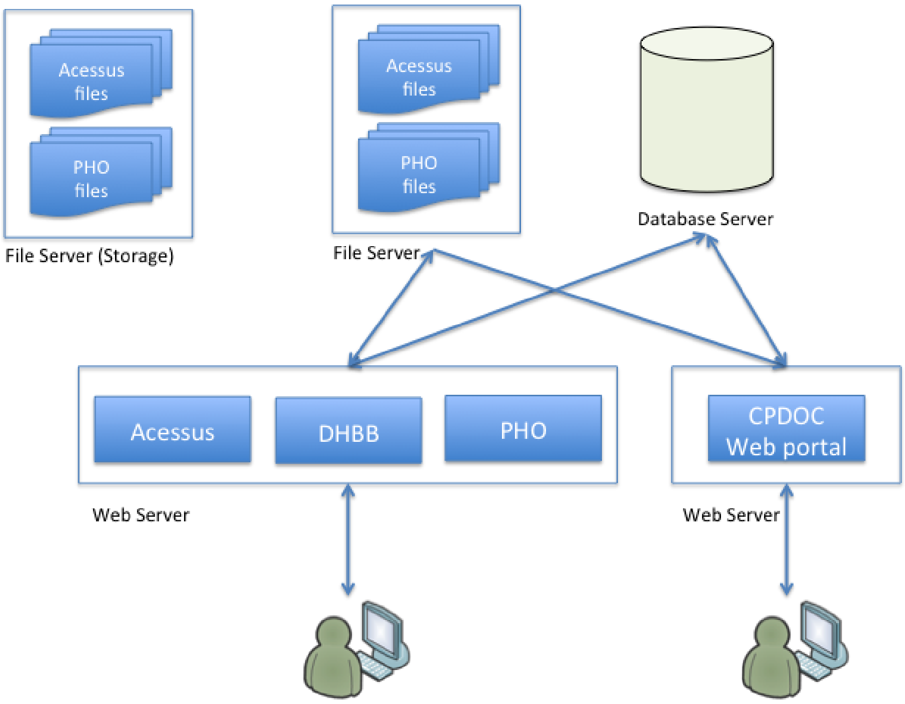


Fig. 1 CPDOC's current architecture: Accessus, DHBB and PHO information systems (intranet) feed a database server with metadata. The files are stored in file servers (high resolution in a storage and low resolution images and documents in a file server acessible by a web server (CPDOC's website).

comprehends: funds (or archives), series, subseries, document units, documents and pages. For instance, there is the archive "Azeredo da Silveira" where one of the series is called "Ministry of Foreign Affairs" which in turn has the subserie "Inter-American Affairs". One of the document units of this subseries deals with the theme "Environment", containing various documents such as telegrams, reports, speeches etc. Another document unit in this subseries covers the subject "Nuclear Agreement", for instance.

Together, they comprise nearly 1.8 million documents or 5 million pages. From this, nearly 900 thousand pages are in digital format and it is expected that they will be all digitized in the next few years. The collection entries metadata are stored in the database. It can be accessed through the institution's intranet for data maintenance or via the CPDOC's website for simple data queries. Currently, the queries are restricted to keyword searches linked to specific database fields defined in an *ad hoc* manner. For those documents that are already digitized, two digital file versions were generated: one in high resolution aiming at long-term preservation and another in low resolution for web delivery. High resolution files are stored in a system with disk redundancy and restricted access, while low resolution files are stored in file servers (Figure 1).

2.2 Oral History Interviews (PHO)

CPDOC's collection of Oral History entries hosts currently more than 6.000 hours of recording, corresponding to about 2,000 interviews. More than 90% of those, video or audio, are in digital format. For the time being, two kinds of queries are available for the database: query by subject and query by interviewee. Each interview record holds brief technical information and a textual summary with descriptions of the interview themes in the order they appear in the recording. Almost 80% of the interviews are transcribed, but to access the audio and video content the user needs to come in person to the CPDOC. Currently, the institution is analyzing different aspects such as the best format, use policies, access control and copyright issues for making this data available online. As in the case of Accessus, the database actually stores only the interviews metadata, while the digitized recorded audios and videos are stored as digital files in the file servers (storage system).

The PHO data comprises a set of interviews whose interviewees are chosen according to the project funding the initiative for collecting the data. These projects are usually linked to political events, and therefore the persons interviewed are mainly the ones that took part on them.

2.3 Brazilian Historical-Biographic Dictionary (DHBB)

The Brazilian Historical-Biographic Dictionary (DHBB) is widely considered one of the main research sources for information on contemporary Brazilian politicians and themes. It contains about 7.500 entries of biographic and thematic nature, i.e., people, institutions, organizations and events records carefully selected using criteria that measure the relevance of those to the political history of the given period. The entries are written objectively, trying to avoid, as much as possible ideological or personal judgments. CPDOC researchers carefully revise all entries to ensure accuracy of the information and uniform style.

The DHBB relational model can be summarized as one main table that contains a text field with the dictionary entries text encoded in HTML [41] with a set of auxiliary tables that provide keys for metadata values such as: professions, governments, places etc. The current dictionary entries are created and revised in text editors outside the system and are imported into the database. DHBB's database stores very few metadata concerning each entry, the amount of metadata differ between recent entries and old ones. The available queries are limited to keyword searches of the title or the text of the entries.

3 Issues and Opportunities

CPDOC's archives are maintained by three different information systems based on traditional relational data models. This infrastructure is hard to maintain, improve and refine, and the information they contain is not found by standard search engines for two main reasons: (1) the CPDOC website HTML pages are created dynamically only after a specific query is issued; (2) users are required to login to the CPDOC's website in order to issue queries or access the digital files. Service providers do not reach the data directly and therefore cannot provide specialized applications using it. Users themselves are not able to expand the queries over the collections, being limited to the available search interface. In summary, data in CPDOC's collections can be considered to be currently limited to the so called "Deep Web" [3].

CPDOC's systems maintenance is difficult and improvements are hard to implement and therefore innovative initiatives are hardly ever adopted. A relational database model is not easily modified since it is supposed to be defined *a priori*, i.e., before data acquisition. Moreover, changes in the database usually require changes in system interfaces and reports. The whole workflow is expensive, time consuming and demands professionals with different skills from interface developers to database administrators. For instance, in the current data model, any enrichment of DHBB entries with metadata extracted from natural language processing of the entries texts would require a complete adaptation of the relational model, new tables and columns would need to be added and the current SCRUD [2] interfaces would needed to be adapted.

CPDOC's collections do not follow any metadata standards, which hinders considerably the interoperability with other digital sources. Besides, the available queries usually face idiosyncratic indexing problems with low rates of recall and precision. These problems are basically linked to the *ad hoc* indexing strategy adopted earlier to define the database tables and fields.

Finally, data storage is also an issue. Digitized Accessus documents and Oral History interviews are not stored in a single place, but scattered into different file systems and servers. The database only stores the metadata and file paths to the file servers, making it very difficult to ensure consistency between files, metadata information and access control policies.

4 Requirements

The requirements for the migration of this whole framework have different perspectives: the users, scholars, students and researchers within the FGV or other institutions; developers and IT specialists; curators and administrators of CPDOC collections. In this section, we list the ones we identified.

From the users perspective, more flexible ways to interact with the data are needed. Non-technical users may want to use web query interfaces complemented with faceted results, but more advanced users expect to be able to make more flexible queries, possibly exploring relations to entities in other datasets. Another class of users that we aim to support are service providers. That is, developers or companies interested in using CPDOC datasets for novel applications such as the creation of learning objects [3] or systems for online courses like MOOC courses [4]. Such advanced users or developers need to have direct access to the data.

Internal developers, that is, the technical staff working for CPDOC, need to answer the demands of new requirements made by the historians and CPDOC researchers. For these goals, the reuse of tools and modeling decisions made by the community of historians and archives' curators play an important role. The adoption of standard vocabularies and open source systems for digital content management, version control, website generators and search engines can improve considerably their response time of new features and interfaces.

[2] This is an acronym for specifying information systems that usually implement the search, create, read, update and delete operations, `http://goo.gl/33piYJ`.

[3] `http://en.wikipedia.org/wiki/Learning_object`.

[4] `http://en.wikipedia.org/wiki/Massive_open_online_course`.

Researchers and historians of CPDOC, the data curators, need an agile and flexible workflow for adoption of inovations. This means that new features should be easier to be tested and implemented. Improving the data model with new properties or entry types should be as painless as possible. It should not demand too much effort for the adaptation of the existing systems and transformation of the already available data. The curators also desire more interoperability with other well-known datasets, ontologies and vocabularies like DBPedia [6], GeoNames [45], YAGO [42], SUMO [35] etc. Such interoperability can, promote and improve the publication of the CPDOC archives. Contributions from the community in the improvement of the quality and volume of the collections is also desired. In this sense, interoperability can help in the engagement of the community once they perceive the value, transparency and availability of the data. However, these contributions should be curated and versioned since the trust and quality of the collections and their metadata is an important asset of CPDOC.

5 The suggested architecture

Relational databases are often hard to maintain and share. Moreover, the idea of having in-house developed information systems is being increasingly replaced by the concept of open source systems. In such systems the updating and creating of new features is not sustained by a single institution but usually by a whole community that shares knowledge and interests with associates. In this way the system is kept up-to-date, accessible and improving much faster due to the increased number of contributors. Such systems are also usually compatible with standards so as to ensure they can be widely used.

The intention is to equip CPDOC with modern open data tools so that the way data is maintained, stored and shared can be improved. The proposal focuses on open source systems as a lightweight and shared way of dealing with data. More concretely, it is proposed the replacement of the three CPDOC information systems by the technologies described in the following paragraphs.

Concerning PHO and Accessus data management demands, we believe that a digital repository management system (DRMS for short) would be suitable for the task at hand. DRMSs have all desirable features that are not found in Accessus or PHO, such as: (1) flexible data model based on standard vocabularies like Dublin Core [25] and SKOS [27]; (2) long-term data preservation functionalities like tracking and notifications of changes in files; (3) fine-grained access control policies; (4) flexible user interface for basic and advanced queries; (5) compliance with standard protocols for repositories synchronization and interoperability (e.g. OAI-PMH [28]); (6) import and export functionalities using standard file formats and protocols. In our proposal the metadata and files from Accessus and PHO systems are to be stored in Dspace [5], but any other popular open source institutional repository software such as Fedora Commons Framework [6] would be equally suitable.

With respect to the DHBB, the nature of its data suggests that its entries could be easily maintained as text files using a lightweight human-readable markup syntax. The files can be organized in an intuitive directory structure and kept under version control for structured and collaborative maintenance. The use of text files can be justified by: (1) easiness of maintenance using any text editor (tool independence); (2) conformity to long-term standards by being software and platform independent; (3) easiness to be kept under version control by any modern version control system [7] since they are textually comparable; and (4) efficiency of information storage. [8]

The adoption of a version control system will improve considerably the current workflow of DHBB reviewers and coordinators. Today the workflow is basically composed by the creation of entries using Microsoft Word and the exchange of emails. The adoption of the new tool will allow file changes to be tracked, implementing a process of collaborative creation without the need of sophisticated workflow systems. This follows the methodology developed by open sources communities for open source software maintenance. Git [9] is specially suited since it ensures data consistency and keeps track of changes and authorship in a collaborative development environment.

5.1 Migration Schema

Many of the described proposals are already implemented as a proof of concept prototype to evaluate the viability of this environment in CPDOC. Figure 2 illustrates the necessary steps to fully implement the project. In the following paragraphs we briefly describe these steps.

Step (1) is already implemented: the relational database was exported to RDF [32] using the open source D2RQ [5] tool. The D2RQ mapping language [14] allows the definition of a detailed mapping that implements the migration of a given relational model to a graph model based on RDF, as sketched out in [4]. The mapping created so far defers any model improvement to step (2) described below.

Step (2) represents a refinement of the graph data model produced in step (1). The idea is to create a data model based on standard vocabularies like Dublin Core [25], SKOS [27],

[5] `http://www.dspace.org/`

[6] `http://www.fedora-commons.org`

[7] `https://en.wikipedia.org/wiki/Revision_control`.

[8] In text files, all DHBB entries uses less than 50% of the current space used by the entries saved as HTML in the database.

[9] `http://git-scm.com`

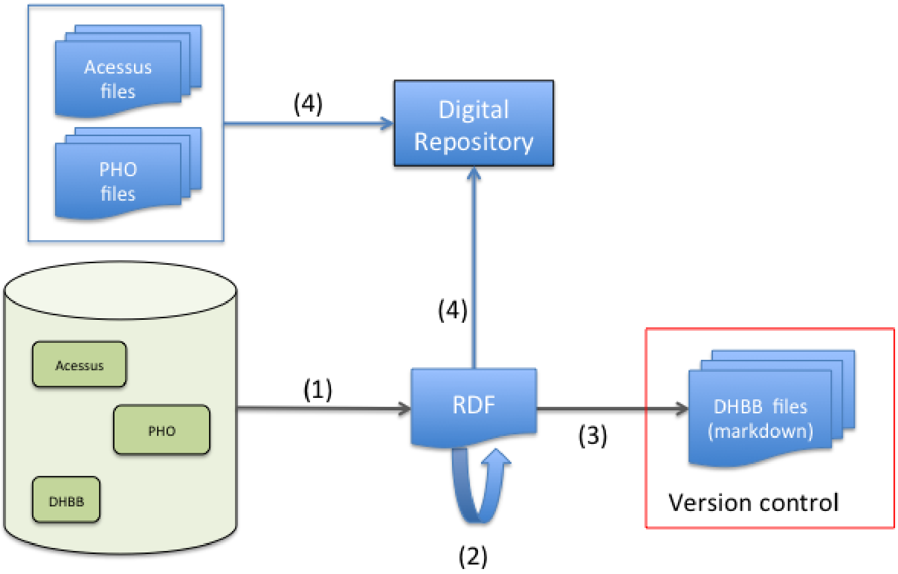


Fig. 2 Migration schema from current relational databases model to the proposed model. Acessus and PHO files are to be stored in a digital repository, while DHBB files are version controlled using a web hosting service. The metadata is converted from relational databases to a RDF database.

PROV [21] and FOAF [8] using well-known conceptual models like [12]. The adoption of standard vocabularies makes the data interchangeable with other collections and facilitates its use by service providers and other users. In Section 6 we describe a refinement proposal to also complete and extend the available metadata using tools to process the raw data available in different file formats such as images, sound and text.

In Step (3), which is also implemented, we deploy a text file for each DHBB entry. The files use YAML [2] and Markdown [22] markup languages to describe the metadata and the entry content. YAML and Markdown were chosen because they are both text-based, human-readable and are supported by nearly all static website generators.

In the planned step (4) digital files and their metadata will be stored in a DRMS. This step is more easily implemented using the RDF produced in step (2) than having to access the original database for two main reasons: SPARQL `CONSTRUCT` queries allow the extraction of graph patterns from RDF into RDF subgraphs; and most of DRMS systems maintain metadata of items already in RDF format.

Considering that all DRMS have detailed control access mecanisms, both high and low resolution files can be imported to the same digital repository making only the low resolution open for public access. This is necessary mainly to preserve bandwidth and because in general, the low resolution files also contain some watermark and embeded metadata.

The proposed architecture for CPDOC's archives maintenance is presented in Figure 3. We emphasize that one of our main goals is to make the collections available as open linked data. This can be accomplished by releasing data as RDF and OWL files for download or by providing a SPARQL Endpoint [10] for queries. Since data evolve constantly, the team of CPDOC would deliver periodical data releases and updates. Apart from the RDF and OWL files and the SPARQL Endpoint, it is also important to provide a lightweight and flexible web interface for final users to browse and query data. This can be done using a static website generator and Apache Solr [10] for advanced queries. As a modern index solution, Solr can provide more powerful and fast queries support when compared to traditional relational database systems. The use of a static site generator [11] allows the maintainers to have full control over the release of new data on the web. It is interesting to notice that CPDOC's maintenance workflow fits well the static website generator approach, since the generated website only needs to be updated, manually, when a new collection is scheduled to be published.

In addition, this approach allows the generation of stable URLs for each relevant entry. For instance, each DHBB entry can have a stable URL that can be indexed by standard search engines. In this way different and complementary outputs are delivered for different purposes and users: a website for browsing, RDF and OWL files for downloading and SPARQL Endpoints for queries.

The results obtained so far encouraged us to propose a complete data model aligned with open linked data vocab-

[10] `http://lucene.apache.org/solr/`

[11] In this application we used Jekyll, `http://jekyllrb.com`, but any other static site generator could be used.

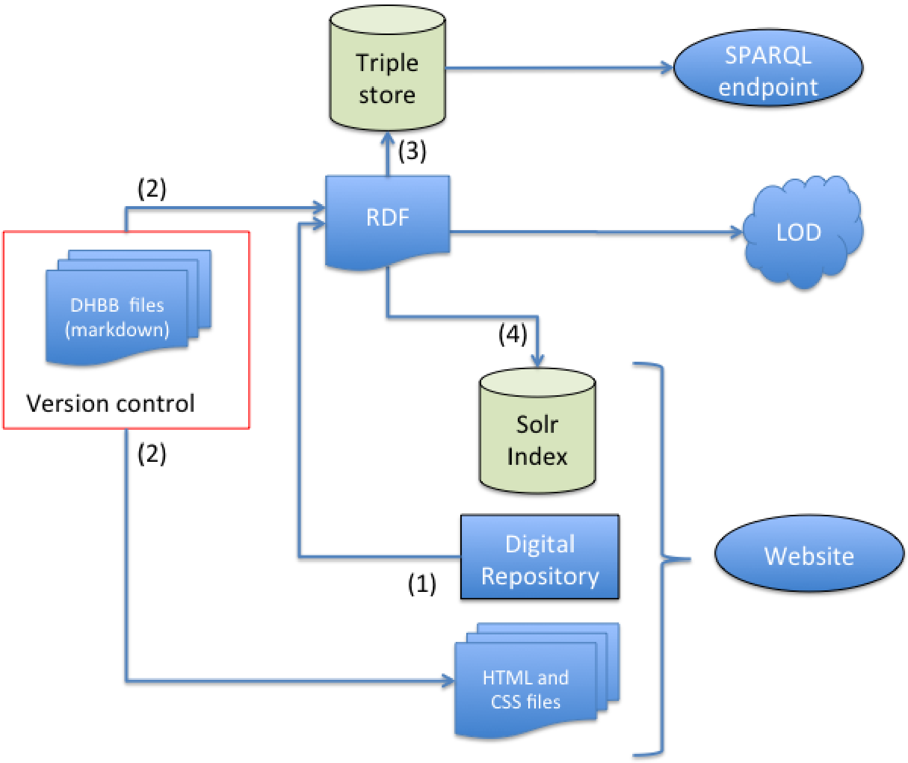


Fig. 3 The final architecture stores the data in digital repositories and web hosting services and the metadata in a RDF database. The data is accessible by different interfaces: a SPARQL endpoint using a triple store; a website providing query tools; or directly making the RDF available to service providers using Linked Open Data.

ularies and enhanced by pattern recognition techniques, as presented in detail in next section.

6 Enhancing the Data

More than simply improving the current infrastructure for storing and accessing data, we would like to exploit the possibilities of CPDOC's archives as sources of knowledge. We discuss in this section four ways of enhancing the data in the archives, using complementary technologies.

6.1 Semantic Web technologies

One of the possible ways of enhancing information extraction from the CPDOC archives is to embed knowledge from other information sources by creating links among the available data. Since much of the data in the archives is related to people and historical events some pre-existing and available ontologies and vocabularies can be used in this task. The nature of the data allows us to use projects that are already well developed for describing relationships and bonds between people, such as FOAF [8] (Friend of a Friend) – a vocabulary which uses RDF to describe relationships between people and other people or things. FOAF permits intelligent agents to make sense of the thousands of connections people have with each other, their belongings and historical positions during life.

A second example is the use of PROV [21] vocabulary, created to represent and exchange provenance information. This is useful to gather information about data that can be structurally hidden in tables or tuples. The RDF graph model also enables the merging of data content naturally. The DBpedia project, for instance, allows users to query relationships and properties associated with Wikipedia resources. Users can link other datasets to the DBpedia dataset in order to create a big and linked knowledge base. CPDOC could link their data to DBpedia making it available to a much larger audience.

We now discuss an example of enrichment of a fragment of CPDOC's data. Figure 4 shows a fragment of the current RDF model produced by D2RQ (in step (1) of Figure 2) using the original CPDOC database relational model. This fragment shows only some PHO classes (derived from the tables) and some properties (derived from the foreign keys). Classes are written inside the boxes and properties are represented by the names in arrows that connect boxes.

The model presented in Figure 4 corresponds to the raw D2RQ results and shows that D2RQ was not able to automatically improve much the existing model. D2RQ was

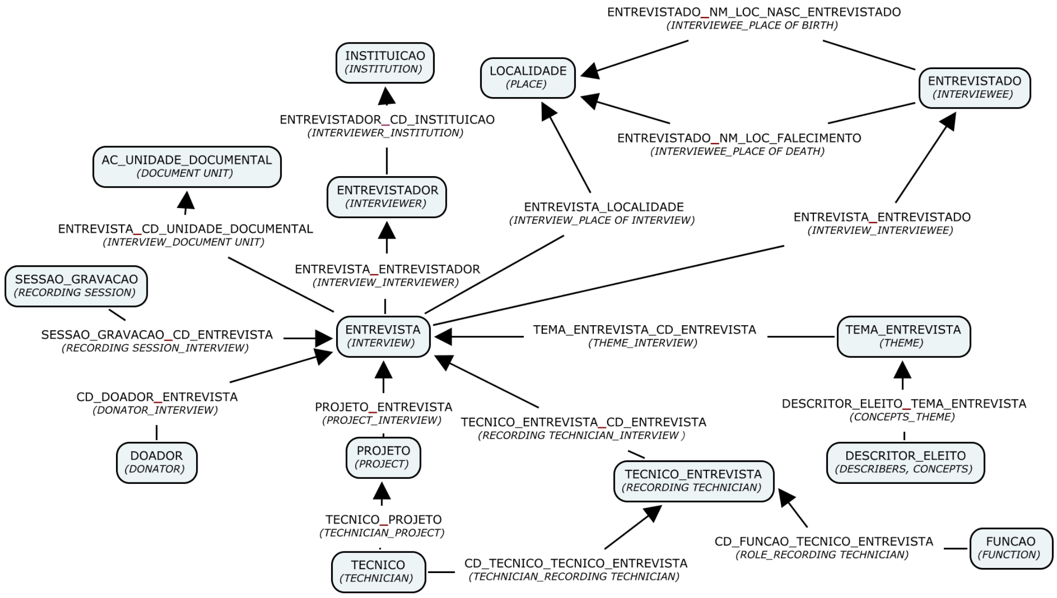


Fig. 4 RDF model example from PHO using the D2RQ conversion tool. This result is to be refined.

able to correctly translate N:M relationships in the relational model, such as `entrevista_entrevistador` (interview/interviewer) (originally a table in the relational model) to a property that connect directly instances of `entrevista` (interview) to instances of `entrevistador` (interviewer). Nevertheless, the N:M relationship between `entrevista` (interview) and `tecnico` (technician) was kept in the intermediary class `tecnico_entrevista` (technician/interview) due to the existence of additional information: the role of the interview technician (class `funcao`). The relational model also seems to have some inconsistencies.

For instance, although the connection between technician and interview is parametrized by different roles, the donor, interviewer and interviewee of a given interview are represented each one in a specific table. Moreover, interviewee, interviewer, donor and technician are all people, and as so they share common properties like name, address etc, and therefore could be modeled using a *person* class (more specifically, the `foaf:Person` from `FOAF` vocabulary), for instance.

Figure 5 shows how the PHO model can be refined in our approach. The new model uses standard vocabularies and ontologies, making the whole model much more understandable and interoperable. The activity box describing provenance `prov:Activity` was duplicated just for easiness of presentation. The prefixes in the names indicate the vocabularies and ontologies used: `prov` [21], `skos` [27], `dcterms` and `dc` [25], `geo` [46], and `bio` [15]. We also defined a CPDOC ontology that declares its own classes and specific ontology links, such as the one that states that a `foaf:Agent` is also a `prov:Agent`. In Figure 5, we see that some classes can be subclasses of standard classes (e.g. `Interview` is a `prov:Activity`), while some others can be replaced by standard classes (e.g. `LOCALIDADE` (location) by `geo:Place`).

The main advantage of adopting well-know vocabularies is that users and researchers from other institutions are able to understand and use CPDOC data. Moreover, by using vocabularies like FOAF and PROV, adopted by many other data providers, we improve accessibility, that is, the possibility of links between entities from CPDOC's data and entities residents in data from other data providers. This network of interlinked entities ultimately will generate more knowledge from the available data.

6.2 Lexical Resources and Natural Language Processing

Another way of enhancing data from historical archives is by means of Natural Language Processing (NLP) methods such as question answering. We would like, for instance, to be able to answer generic questions about the entries in the DHBB database, such as "to which top 5 schools most of the Brazilian leaders of the beginning of the 20th century went to?"

To answer generic questions as well as for many other knowledge intensive tasks, the use of lexical resources such as WordNet [18] is indispensable. It is well-known that WordNet is an extremely valuable resource for research in Computational Linguistics and NLP in general. WordNet has been used for a number of different purposes in information systems, including word sense disambiguation, information retrieval, text classification and summarization, and dozens of other tasks.

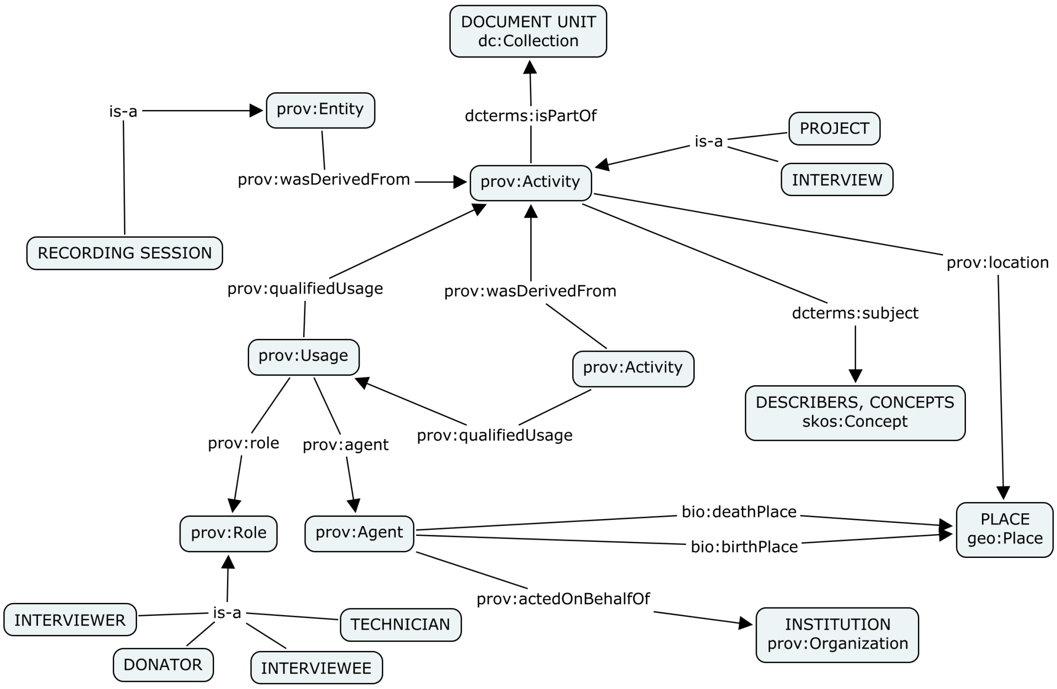


Fig. 5 PHO revised RDF model. Using some of the commonly used standards for Linked Open Data it is possible to make the data more accessible and semantically meaningful.

Given that the texts in CPDOC's archives are written in Brazilian Portuguese, it is convenient to be able to use a Brazilian version of Wordnet, such as the OpenWordnet-PT [38], developed by some of the authors, to address NLP tasks. OpenWordnet-PT is being developed with the support of FGV. The goal of that project, in the long run, is to use formal logic tools to reason about knowledge obtained from texts in Portuguese. OpenWordnet-PT is available for download [12] and query [13] in the Open Multilingual Wordnet website [7].

OpenWordnet-PT is being improved by drawing on a two-tiered methodology that offers high precision for the more salient and frequent words of the language, but also high recall to cover a wide range of words in the desired corpora. We combined manual base concepts annotation with statistical cross-lingual projection techniques to obtain the first version of the resource. Recently, we combined OpenWordnet-PT with NomLex-PT [11]. We started Nomlex-PT with a manual translation of the original English NOMLEX [31] to Brazilian Portuguese. Incorporating NomLex-PT data into OpenWordnet-PT has shown itself useful in pinpointing some issues with the coherence and richness of OpenWordnet-PT.

For the time being, we have mainly used the DHBB data as a way of checking the coverage of nominalizations in NomLex-PT and OpenWordnet-PT. The DHBB corpus is very well suited to observe nominalizations in Portuguese, since it is somewhat erudite, written in higher register than newswire, but meant to be accessible to students, and domain specific: historical data lends itself to conceptualizations that are usually expressed via nominalizations. In one small, but telling experiment, we used Freeling [36] to automatically process the entry data of the DHBB by performing tokenization, sentence splitting, part-of-speech tagging, and word sense disambiguation with respect to OpenWordnet-PT. Then we manually checked the nouns (over a certain threshold) that were nominalizations and verified how many of these were already included in both resources: NomLex-PT and OpenWordnet-PT. This gave us confidence that the coverage of OpenWordnet-PT is reasonably good for nouns (specifically for nominalizations), as well as providing a collection of "concepts" that should be the seed for a DHBB-History Ontology, which we hope to develop next. Named entities recognition and other NLP tasks can automatically create connections that improve dramatically the usability of the content of the DHBB. Resources like YAGO [42] and BabelNet [33] link Wikipedia to WordNet. The result is an "encyclopedic dictionary" that provides concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations. The SUMO Ontology [35] could also be used to provide a complete formal definition of terms linked to WordNet, over which we can do automated inference, using theorem provers and proof assistants.

Until now, DHBB entries have been used as a clean and high-quality corpus that is helping us to improve the lexicon in OpenWordnet-PT. Much more needs to be implemented,

[12] `https://github.com/arademaker/openWordnet-PT`

[13] `http://logics.emap.fgv.br/wn/`

such as, the similar work done by [19] for information extraction.

6.3 Audio alignment technologies

Another very promising approach for enriching CPDOC's data is to exploit the audio and video interviews stored by PHO (Section 2).

Currently about 75% of PHO interviews are transcribed. The transcription process is manual and provides a fluid and correct text, often omitting disfluencies, irrelevant grammatical errors or hesitations. The NLP techniques and resources from Section 6.2 are clearly suited for handling transcriptions, but the alignment of audio and the corresponding transcription plays an important role to make audio data available within the semantic structure hereby proposed.

The manual alignment of audio and transcription is tedious, very time consuming and therefore virtually unmanageable even for a small amount of data. We propose the use of a couple of open-source tools to automatically align PHO audio and transcription files. The first one is the Hidden Markov Model Toolkit (HTK) developed by the University of Cambridge [47]. This tool is specially suitable for speech recognition tasks based on Markov models. We also use the collection of tools provided by the FalaBrasil Research Group at the Federal University of Pará (UFPA): Portuguese acoustic model, Portuguese language model, phonetic dictionary and grapheme-phoneme converter [34].

In order to compare words in text and audio, we identify the phonemes present in a given transcription, but the sequence of characters that compose words hardly ever has a straightforward phonetical translation. The same character corresponds to different sounds (phonemes) depending on the word in which it is included (e.g. the character "u" in the words "fun" and "full"). Therefore, a new text file is generated from the transcription containing the words expressed by means of their phonemes, to allow the comparison between text and audio. The "UFPAdic3.0" phonetical dictionary is used to create this text file of phonemes, and contains about 64800 words in Portuguese and the corresponding phonemes. The phonemes of words not present in the dictionary are estimated using the "Conversor Grafemafone 1.6" grapheme-phoneme converter, also provided by the group FalaBrasil.

The most computationally intensive step of the alignment procedure is the matching between the phonemes from the transcription and the audio file. For this alignment we use an acoustic model (LapSAM v1.5) for Portuguese based on Markov chains provided by the FalaBrasil group with 16 hours of recording, in a format compatible with HTK. This model allows the estimation of the phonemes more likely to have been pronounced at a given audio snippet. The process consists of maximizing the global likelihood of matching the phonemes estimated from the transcription and the ones present in the audio sequence.

To perform this optimization, we used the Viterbi algorithm, also available in HTK. Instead of using the algorithm for the whole audio content, we decided to split the audio in smaller snippets and process them sequentially. This was done after we had problems to processing the whole data at once, which delivered inaccurate results for audio over 10 minutes long. This is probably due to the very nature of Viterbi algorithm, which performs a global optimization, and therefore might deliver inaccurate and computational time demanding results for long audio snippets. To tackle this problem, we split the audio in overlapping snippets of 5 minutes, processing the snippets and taking only the first minute of each snippet in the final result. This simple procedure provides a more accurate alignment within the snippet, speeds up the process, and guarantees that long interviews can be handled properly.

The alignment procedure outcome is a file with timestamps for each word in the audio within the given transcription. These words are grouped to create timestamps for sentences with 30 to 45 characters long and ultimately organized as a subtitles file. Using this file we are able to construct an user interface that allows users to query for a given word in the transcription text and be automatically redirected to the time in the audio when the queried word is spoken.

The alignment of audio and transcription allows the semantic linking of the audio data in the interviews to other textual and audiovisual data, from Wikipedia or DHBB, for instance. This improves not only the data present in the audio files, but also the external data that can be linked to them, integrating CPDOC archives in their different supports.

6.4 Image processing technologies

Historical images play an important role in CPDOC collections. They hold much information and usually represents a challenge to extract semantically the information for connecting visual with textual data. The Accessus system stores historical photos and their metadata, which is expected to explain part of the information the image holds. Still, image processing techniques can help to extract unknown information from images, or tagging information graphically into the photos themselves.

Over the last decade several photographic collections have been digitised and many of them have been made available for public access through web portals, for instance see the Library of Congress' photo-stream on Flickr [30]. Each image may potentially have captions and/or texts that have been produced by experts to describe its content, which is usually stored as free text within a data basis. Captions may refer to the picture as a whole and/or describe a specific important feature that occurs in a particular subregion of the image. In order to specify the referred subregion with natural language, sentences like: *on top of*, *on the right of*, *in the first plane*, *from right to left*, *dressed in white*, *using a red hat*, etc. are used frequently. All of these sentences suffer from lack of precision, from ambiguity, and their automatic processing can be difficult. Nowadays, the information retrieval of structured information is appealing and the migration of natural language captions to structured information is desirable in a variety of photographic collections.

The CPDOC photographic archive has been arranged and handled manually in its organizational phase. In 2008, an extensive digitization project began, where the images and the results of the intellectual process of character identification and captioning were made available for public access through a web information portal. However, with the evolution of multimedia collection retrieval resources introduced by the use of semantic standards, the need to convert the collection to semantic standards arose.

The CPDOC's photographic collection has important idiosyncrasies for image processing: (1) Non-frontal faces appear very frequently (non-trivial for detection and recognition); (2) typically characters that are important within a single photo are just a few when compared to the number of faces that actually appear in the image; and (3) many of the images present some characteristic that makes the automatic image processing pipeline harder than usual; more specifically, they may be monochromatic, contain different types of noise and may present the characters in very low resolution. These characteristics lead us to discard the use of off-the-shelf character annotation tools and libraries, such as Google's Picasa photo organizer and editing software [14]. Of course, one additional reason to discard the adoption of any off-the-shelf tool is the need of a integrated system in the CPDOC's workflow of image annotation and archiving. But the most evident limitation of the majority of the available photo annotation tools and libraries is that they were not designed to process information available in captions or texts produced by experts that describe the content of previously organised photographic collections.

To address our specific problems we developed the VIF (Very Important Faces) software [44] as an environment for describing image contents throughout analysis, annotation and verification of multi-modal metadata (see Figure 6). The VIF combines face detection techniques and basic regular expressions [15] to help the user in the association of names that occur in the legend to faces that occur within an image.

In order to achieve the desired result, faces of important historical figures need to be detected within the images. The *face detection* task consists of identifying subregions within an image where a human face occurs. Face detection is a well developed research subject and is already an off-the-shelf technique for the frontal face case. However, its general form is still a challenging computer vision problem due to variations in head pose orientation, facial expression, occlusions, variation in lighting and imaging conditions and the presence of non-uniform backgrounds.

Concerning the textual information present in the captions, an automatic proper name extraction task was implemented in order to ease face tagging. We took advantage of the description dictionary provided by Accessus to simplify the proper name extraction task using regular expressions.

The expected outcome consists of descriptions of the faces' spatial positions within the images, followed by the name matched to the face. We call attention to the fact that the face's spatial position annotation is more precise than the natural language annotation and can potentially solve some of the ambiguities present in captions; the drawback is that spatial descriptors are easily readable by machines, but not by people. Thus we generate the demand of a layer of processing between the information and the user.

In order to support the experts to efficiently review the annotation produced within VIF, which aims to avoid error inclusion in the database re-annotated by non-experts or by automatic processing, we proposed the adoption of a set of annotation maturity levels to tag annotation provenance. Image embedded metadata seems to be the natural tendency for dealing with questions related to association of provenance of the image content description. The role of an expert is to guarantee a high level of confidence in the information associated with the photo, but experts can be expensive. Therefore, it is desirable to use this resource efficiently. Considering databases that have been previously annotated by experts, the migration of this information to structured standards would become infeasible if experts were required to redo the annotation task. Thus, less costly solutions should be proposed, that is, the task has to be performed by non-experts, crowd-sourcing or automated means, leaving to the expert the task of verifying the annotation.

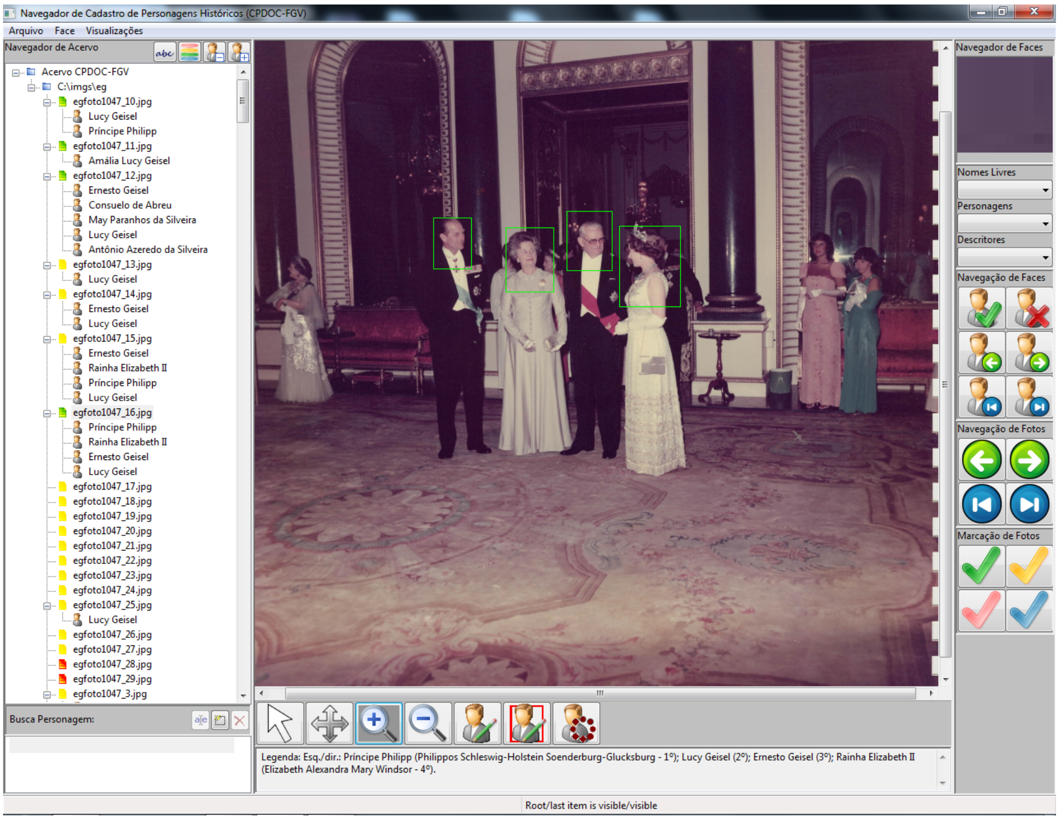
VIF has been designed to attend the requirements of a contemporary history multimedia database, that is, considering the specific demands of archivists that wish to annotate the occurrence of important characters in photographic collections. It helps to extract information (or to link existing metadata) from images, allowing to semantically create connections with the other CPDOC's collections.

7 Evaluation Plans

The initiatives discussed here will provide easy and intuitive tools for researchers to access the CPDOC historical data, maintained, via their semantics, independently of the

[14] `http://picasa.google.com`

[15] `http://en.wikipedia.org/wiki/Regular_expression`.



- Acervo CPDOC-FGV
  - C:\imgs\eg
    - egfoto1047_10.jpg
      - Lucy Geisel
      - Príncipe Philipp
    - egfoto1047_11.jpg
      - Amália Lucy Geisel
    - egfoto1047_12.jpg
      - Ernesto Geisel
      - Consuelo de Abreu
      - May Paranhos da Silveira
      - Lucy Geisel
      - Antônio Azeredo da Silveira
    - egfoto1047_13.jpg
      - Lucy Geisel
    - egfoto1047_14.jpg
      - Ernesto Geisel
      - Lucy Geisel
    - egfoto1047_15.jpg
      - Ernesto Geisel
      - Rainha Elizabeth II
      - Príncipe Philipp
      - Lucy Geisel
    - egfoto1047_16.jpg
      - Príncipe Philipp
      - Rainha Elizabeth II
      - Ernesto Geisel
      - Lucy Geisel
    - egfoto1047_17.jpg
    - egfoto1047_18.jpg
    - egfoto1047_19.jpg
    - egfoto1047_20.jpg
    - egfoto1047_21.jpg
    - egfoto1047_22.jpg
    - egfoto1047_23.jpg
    - egfoto1047_24.jpg
    - egfoto1047_25.jpg
      - Lucy Geisel
    - egfoto1047_26.jpg
    - egfoto1047_27.jpg
    - egfoto1047_28.jpg
    - egfoto1047_29.jpg
    - egfoto1047_3.jpg

Legenda: Esq./dir.: Príncipe Philipp (Philippos Schleswig-Holstein Soenderburg-Glucksburg - 1º); Lucy Geisel (2º); Ernesto Geisel (3º); Rainha Elizabeth II (Elizabeth Alexandra Mary Windsor - 4º).

Fig. 6 The VIF interface: the annotation and verification modulus are integrated with colours associated with both photos and faces tags, and also interface buttons. Therefore, users access and edit multi-modal content in a transparent manner.

medium in which the information is stored. Ultimately, our goals are: (1) allow efficient querying over all CPDOC collections; (2) make our data available following the linked data principles [16].

Is is important to stress the main contrast between the new architecture and the current one. In the current CPDOC architecture, data is stored in relational databases and maintained by information systems. This means that any data modification or insertion is available in real time for CPDOC website users. However, this architecture has many drawbacks as mentioned in Section 3, and also the nature of the data does not require continuous updates, which means that the cost of this synchronous *modus operandi* is not needed. Usually, CPDOC teams work on individual projects and therefore new collections, documents and metadata revisions are not released very often.

While most of the suggested technologies are of daily use to people with technical backgrounds, such as software developers, they are not very familiar to people with non-technical profiles. In this context, a big challenge of this approach is to motivate the internal users of CPDOC systems, i.e., these archives maintainers, to invest time to learn new technologies, instead of keeping their inefficient but well-known way of dealing with data.

Our evaluation plans for the new architecture are manifold and will generate, once implemented, qualitative and quantitative insights of the quality leap we expect to see. First, we plan to do surveys to evaluate the user experience when using the new tools proposed in this paper. This survey will give us inputs to statistically quantify how users are affected by the solutions we propose. These statistics will be shared with the community.

Another simple but powerful metric that is going to be analyzed is the evolution of the number of visitors in the query tools and data dumps. We expect that easier and more efficient tools will increase substantially the number of visitors to the CPDOC archives. Since the archives will be made available as RDF following linked data principals, we can quantify the number of dereferences of uris of our datasets made to our server.

We also expect a growth in the involvement of researchers of CPDOC in feeding, analyzing and using the data hosted by CPDOC, in their own individual research agendas. This growth can be verified by the publications and citations made about CPDOC's data and hopefully will indicate an improvement of accessibility and availability of the CPDOC archives for its researchers. We also expect that CPDOC researchers contribute with more insights and ideas for data enrichment given the more flexible architecture.

[16] http://linkeddata.org

However, how to evaluate the results of our efforts for data enrichment? In Section 6, we presented some techniques that we are exploring. At the end of each subsection, we provided some discussion of the expected benefits of each approach. In the next paragraphs, we highlight some aspects of evaluation.

Regarding the image processing for faces detection and annotation, a tricky discussion arose when we proposed to the data curators a semi-automatic annotation approach. Curators considered the automatic annotation, as well as the non-expert annotators, as potentially inserting errors in an information data, which was originally produced by experts. Even if our semi-automatic name to face matching were capable of 95% of correct assignments, that would mean the potential insertions of information with at least 5% of error, which is usually unacceptable. For this reason, it is crucial to keep track of all data produced in automatic or semi-automatic manner. We plan to provide provenance annotations for all automatic or semi-automatic generated data, following the same directions of [19].

When dealing with cultural heritage database annotation, it is crucial to define a maturity level related to the produced annotation. The *annotation maturity level* is to be used, for instance, to decide if a document is ready or not be published. From this discussion, we conceived the VIF annotation verification module. The aim is to support an authorised expert with tools for efficient review of the annotation produced by non-experts, which includes both manual and semi-automatic. We have already made some preliminary evaluation of the VIF usability. The CPDOC's team approved the tool interface and the features. Naturally, the assessment of such tool is not quantitative but qualitative. We are not focused on reducing time spent with annotations nor improving the automatic face recognition or face identification. Our goal is to lessen the necessity of experts for the minimal necessary intervention in the archiving workflow. This kind of evaluation will only be possible with the use of the tool by the CPDOC team for a while.

In order to enhance efficiency in the verification work, VIF offer querying and sorting over both the captions and the annotations. The Interface based on fast information visualisation allows an expert to navigate through quickly and review a summary of the annotation produced by non-experts. In the preliminary experiments that we conducted, we had positive feedbacks by non-experts and experts from CPDOC.

Nevertheless, we can evaluate the identification of important people in historical images in a quantitative manner if necessary. This evaluation can be done by sampling and manually inspecting and using semi-automatic methods to present results for verification. Some findings on this direction were already published [44] concerning this ongoing task.

In [37] we presented the evaluation of natural language processing outcomes on the historical dictionary. The next step regarding NLP for DHBB entries is to experiment with different techniques for information extraction. In the first instance, we are interested in extracting life events (i.e. birth, death, marriage) and professional activities mentioned in the dictionaries entries. We can take advantage of the fact that some entries have already metadata about professional activities to compare the facts extracted from the text with the facts presented in the metadata available.

For evaluating the alignment and automatic transcription of interviews, we plan to sample, and manually inspect the results. Given our limit resources, we also believe that making the data available as soon as possible will allow visitors to provide feedbacks for data improvements.

8 Conclusion

We presented a new architecture for CPDOC's archives creation and maintenance, based on open linked data concepts and on open source methodologies and tools. This effort is expected to have a great impact on the way data is accessed and made available. The identification of entities in historical images, interviews and textual data can make explicit the semantics of these entries and therefore provide interconnections within the archives and with the outside community. This represents a leap of quality of the experience we expect to provide for researchers when consulting the archives. The goal is to provide a smart, multimedia and semantically rich environment, a large knowledge database, accessible using modern standards of semantics.

Among the advantages of the architecture proposed, we highlight that it ensures more control and easiness of data maintenance. It allows easy integration of changes in the data model, without the need for database refactoring. This ultimately means less dependency of the CPDOC's team on the FGV's Information Technology department. We are aware that the migration process of CPDOC's archives from relational database to RDF bases is not a trivial task: we found several sources of data inconsistency and noise, as the data was maintained by many different individuals over a long span of time, using poorly documented protocols. We expect that the staff involved will need to be trained to use the proposed new tools (text editors, version control software systems and command line scripts etc), but this seems worth the trouble, given the benefits outlined.

With regard to automated analysis of the content of the archives, many research opportunities for the open linked use of CPDOC collections were proposed in Section 6. The use of lexical resources in Portuguese is being carried out so as to improve the structure and quality of the DHBB entries and the automatic extension of their mapping can be defined following ideas of [9]. The alignment of audio and transcriptions implemented a way of embedding semantics into audio files of the interviews hosted by PHO; while the CPDOC historical imagery team is benefiting from image processing techniques to automatically extract useful information that can be linked to other collections and accessed more efficiently.

The technologies proposed in this paper for migrating CPDOC's archives to a model with the open linked data perspective can be applied to other digital libraries and archives. Actually they are already being used in many projects cited in the introduction, and can be combined to solve different challenges. In this paper we proposed models for archives using audio, images and textual support. Even though this media are very common, different archives, with different support, probably will have idiosyncrasies that will require specific tools related to the kind of data hosted.

The possibilities offered by the paradigm of open linked data that we intend to bring to CPDOC directly affect the way people collaborate for the construction of knowledge. The knowledge about something is not simply about capturing data. People combine cognitive and perceptual faculties as they interact with the resources that surround them, which means that they create new forms of participating in the construction of knowledge. This paper aligns the CPDOC archives with modern concepts of linked open data, hoping to provide efficient dissemination and creation of knowledge.

References

1. Abreu, A.A., Lattman-Weltman, F., de Paula, C.J.: Dicionário Histórico-Biográfico Brasileiro pós-1930, 3 edn. CPDOC/FGV (2010)
2. Ben-Kiki, O., Evans, C., dot Net, I.: Yaml: Yaml ain't markup language. http://www.yaml.org/spec/1.2/spec.html
3. Bergman, M.K.: White paper: the deep web: surfacing hidden value. journal of electronic publishing **7**(1) (2001)
4. Berners-Lee, T.: Relational databases on the semantic web. Tech. rep., W3C (1998). http://www.w3.org/DesignIssues/RDB-RDF.html
5. Bizer, C., Cyganiak, R.: D2R server-publishing relational databases on the semantic web. In: 5th international Semantic Web conference, p. 26 (2006). http://d2rq.org
6. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia - a crystallization point for the web of data. Web Semantics **7**(3), 154–165 (2009). DOI 10.1016/j.websem.2009.07.002. URL http://dx.doi.org/10.1016/j.websem.2009.07.002
7. Bond, F., Paik, K.: A survey of wordnets and their licenses. In: Proceedings of the 6th Global WordNet Conference (GWC 2012), pp. 64–71. Matsue (2012). http://bit.ly/1aN0Xxd
8. Brickley, D., Miller, L.: FOAF vocabulary specification (2010). http://xmlns.com/foaf/spec/
9. Cafezeiro, I., Haeusler, E.H., Rademaker, A.: Ontology and context. In: IEEE International Conference on Pervasive Computing and Communications. IEEE Computer Society, Los Alamitos, CA, USA (2008). DOI 10.1109/PERCOM.2008.21
10. Clark, K.G., Feigenbaum, L., Torres, E.: SPARQL protocol for RDF. Tech. rep., W3C (2008)
11. Coelho, L.M.R., Rademaker, A., Paiva, V.D., de Melo, G.: Embedding nomlex-br nominalizations into openwn-pt. In: Proceedings of Global WordNet Conference 2014 (2014). To appear
12. Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M.: Definition of the CIDOC conceptual reference model. Tech. Rep. 5.0.4, CIDOC CRM Special Interest Group (SIG) (2011). http://www.cidoc-crm.org/index.html
13. da Cultura, M.: Registro aberto da cultura (r.a.c): Manual do usuário (2013). http://sniic.cultura.gov.br
14. Cyganiak, R., Bizer, C., Garbers, J., Maresch, O., Becker, C.: The D2RQ mapping language. http://d2rq.org/d2rq-language
15. Davis, I., Galbraith, D.: BIO: A vocabulary for biographical information (2011). http://vocab.org/bio/0.1/.html
16. Deborah L. McGuinness, F.v.H. (ed.): OWL 2 Web Ontology Language Document Overview, 2 edn. W3C Recommendation. World Wide Web Consortium (2012)
17. Federal, G.: Governo federal dados abertos (2013). http://dados.gov.br/
18. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA (1998)
19. Fokkens, A., ter Braake, S., Ockeloen, N., Vossen, P., Legêne, S., Schreiber, G.: Biographynet: Methodological issues when nlp supports historical research. In: Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC). Reykjavik, Iceland (2014)
20. Friesen, N., Hill, H.J., Wegener, D., Doerr, M., Stalmann, K.: Semantic-based retrieval of cultural heritage multimedia objects. International Journal of Semantic Computing **06**(03), 315–327 (2012). DOI 10.1142/S1793351X12400107. http://www.worldscientific.com/doi/abs/10.1142/S1793351X12400107
21. Gil, Y., Miles, S.: PROV model primer. Tech. rep., W3C (2013). http://www.w3.org/TR/prov-primer/
22. Gruber, J.: Markdown language. http://daringfireball.net/projects/markdown/
23. Haslhofer, B., Isaac, A.: data.europeana.eu - the europeana linked open data pilot. In: DCMI International Conference on Dublin Core and Metadata Applications. The Hague, The Netherlands (2011). URL http://eprints.cs.univie.ac.at/2919/
24. Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S.: Finnish museums on the semantic web. Journal of Web Semantics **3**, 25 (2005)
25. Initiative, D.C.: Dublin core metadata element set (2012). http://dublincore.org/documents/dces/
26. Initiative, O.D.: Open data initiative (2013). http://www.opendatainitiative.org
27. Isaac, A., Summers, E.: SKOS simple knowledge organization system prime. Tech. rep., W3C (2009). http://www.w3.org/TR/skos-primer/
28. Lagoze, C., de Sompel, H.V., Nelson, M., Warner, S.: The open archives initiative protocol for metadata harvesting (2008). http://www.openarchives.org/OAI/openarchivesprotocol.html
29. LexML: Rede de informação informativa e jurídica (2013). http://www.lexml.gov.br
30. Library of Congress: The library of congress' photostream on flickr. http://www.flickr.com/photos/library_of_congress/
31. Macleod, C., Grishman, R., Meyers, A., Barret, L., Reeves, R.: Nomlex: A lexicon of nominalizations. In: Proceedings of Euralex 1998, pp. 187–193. Liege, Belgium (1998)
32. Manola, F., Miller, E. (eds.): RDF Primer. W3C Recommendation. World Wide Web Consortium (2004). URL http://www.w3.org/TR/rdf-primer/
33. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193**, 217–250 (2012)

34. Neto, N., Patrick, C., Klautau, A., Trancoso, I.: Free tools and resources for brazilian portuguese speech recognition. Journal of the Brazilian Computer Society **17**, 53–68 (2011)
35. Niles, I., Pease, A.: Towards a standard upper ontology. In: Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001, pp. 2–9. ACM (2001)
36. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pp. 23–25. European Language Resources Association (ELRA), Istanbul, Turkey (2012)
37. de Paiva, V., Oliveira, D.A.B., Higuchi, S., Rademaker, A., de Melo, G.: Exploratory information extraction from a historical dictionary. In: Proceedings of 1st Workshop on Digital Humanities and e-Science 2014 (2014). To appear
38. de Paiva, V., Rademaker, A., de Melo, G.: Openwordnet-pt: An open brazilian wordnet for reasoning. In: Proceedings of the 24th International Conference on Computational Linguistics (2012). URL http://hdl.handle.net/10438/10274
39. Purday, J.: Think culture: Europeana.eu from concept to construction. The Electronic Library **27**, 919–937 (2009)
40. Rademaker, A., Higuchi, S., Oliveira, D.A.B.: A linked open data architecture for contemporary historical archives. In: L. Predoiu, A. Mitschick, A. Nurnberger, T. Risse, S. Ross (eds.) Proceedings of 3rd edition of the Semantic Digital Archives Workshop. Valetta, Malta (2013). Workshop website at http://mt.inf.tu-dresden.de/sda2013/. Proceedings athttp://ceur-ws.org/Vol-1091/
41. Raggett, D., Hors, A.L., Jacobs, I.: Html 4.01 specification. Tech. Rep. REC-html401-19991224, W3C (1999). http://www.w3.org/TR/html401/
42. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A Core of Semantic Knowledge. In: 16th international World Wide Web conference (WWW 2007). ACM Press, New York, NY, USA (2007)
43. Szekely, P., Knoblock, C., Yang, F., Zhu, X., Fink, E., Allen, R., Goodlander, G.: Connecting the smithsonian american art museum to the linked data cloud. In: P. Cimiano, O. Corcho, V. Presutti, L. Hollink, S. Rudolph (eds.) The Semantic Web: Semantics and Big Data, *Lecture Notes in Computer Science*, vol. 7882, pp. 593–607. Springer Berlin Heidelberg (2013). DOI 10.1007/978-3-642-38288-8_40. URL http://dx.doi.org/10.1007/978-3-642-38288-8_40
44. Vasconcelos, C.N., Sa, A.M., Carvalho, P.C., Sa, M.I.: Structuring and embedding image captions: the v.i.f. multi-modal system. In: VAST: International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage, pp. 25–32. Eurographics Association, Brighton, UK (2012)
45. Vatant, B., Wick, M.: Geonames ontology (2012). http://www.geonames.org/ontology/documentation.html
46. Wick, M., Vatant, B.: Geonames ontology (2011). http://www.geonames.org/ontology
47. Young, S.J., Evermann, G., Gales, M., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK book version 3.4. Cambridge University Engineering Department (2006)