A linked open data architecture for contemporary historical archives

Alexandre Rademaker[1] Suemi Higuchi[2]
Dário Augusto B. Oliveira[2]

IBM Research and FGV/EMAp

FGV/CPDOC

September 25, 2013

Getulio Vargas Foundation (FGV)

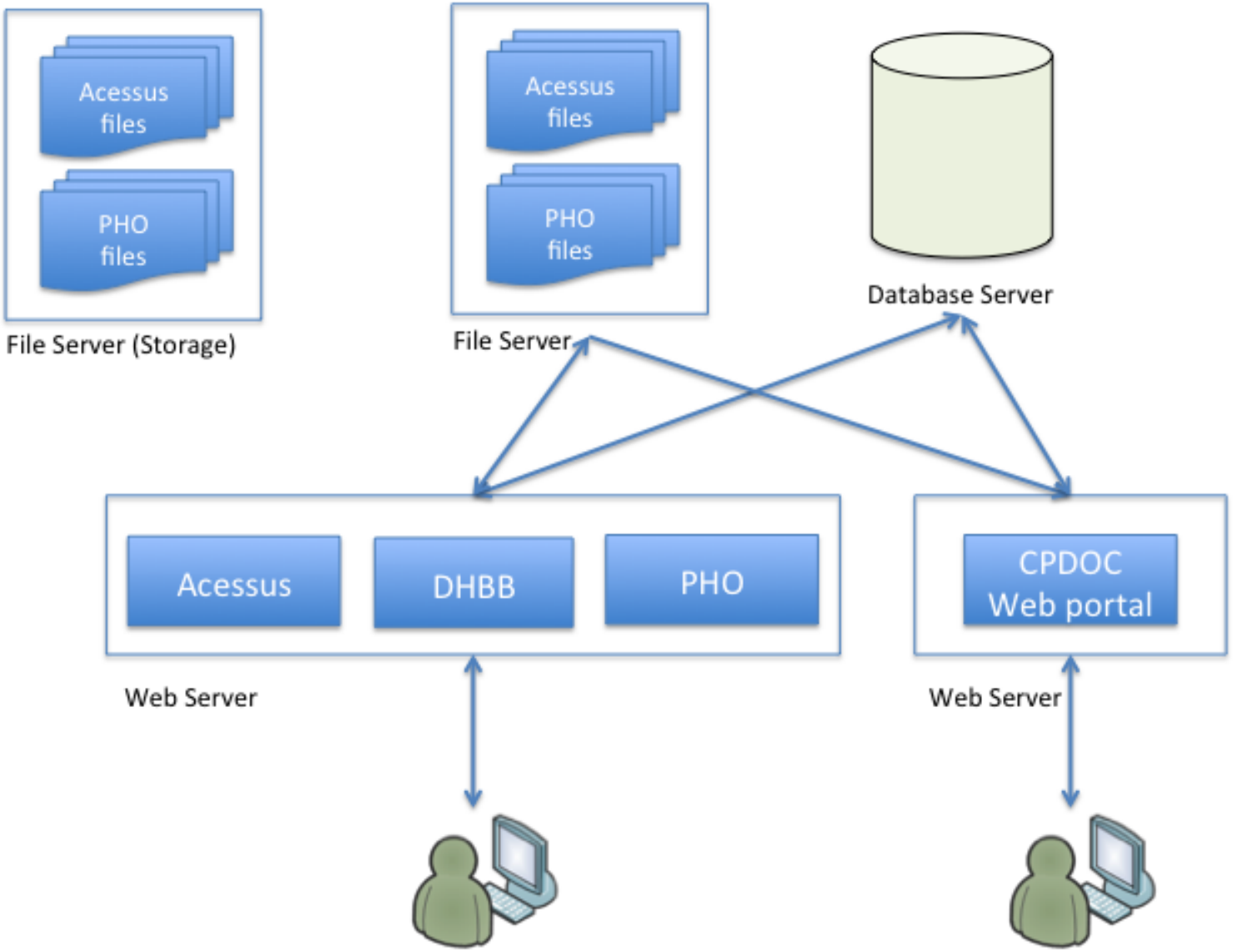
Brazilian higher education and research institution founded in December 20, 1944. It offers regular courses of Economics, Business Administration, Law, Social Sciences and Applied Mathematics. Its original goal was to train people for the country's public- and private-sector management. It is considered by Foreign Policy magazine to be a top-5 policymaker think-tank worldwide.

`http://portal.fgv.br`

CPDOC - Center of Brazilian Contemporary History

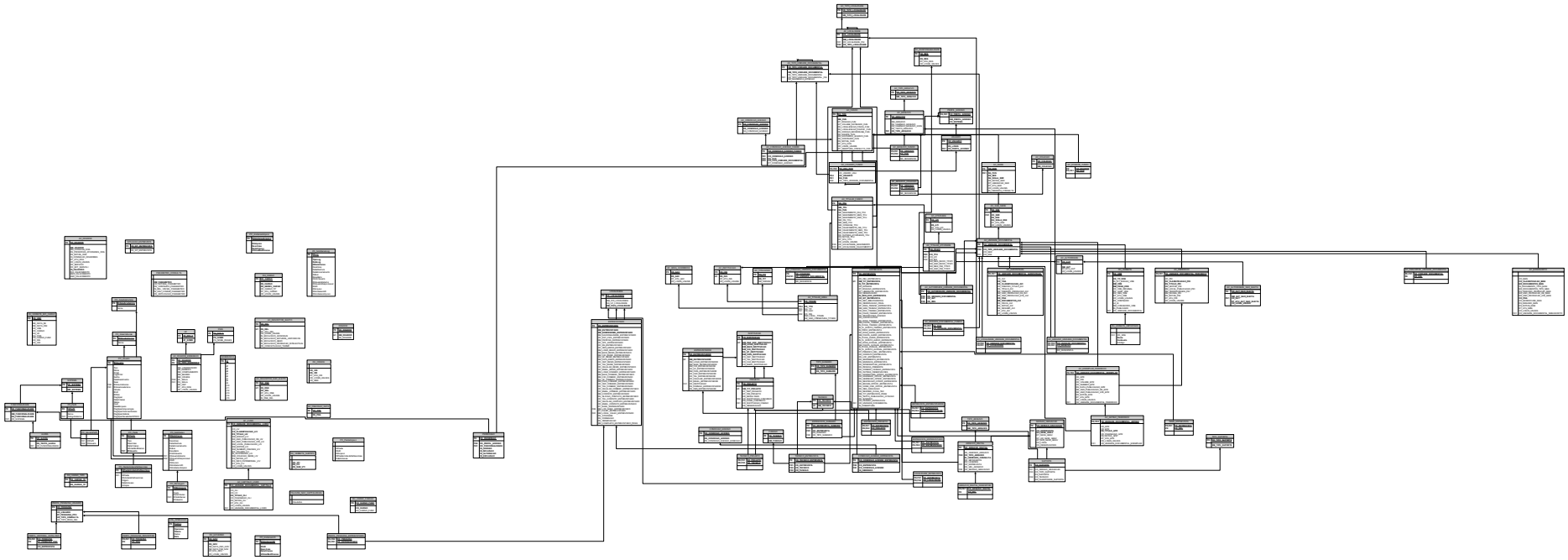
- A major center for teaching and researching in the Social Sciences and Contemporary History located in Rio de Janeiro. It holds:
- **Personal Archives (Acessus)** $\approx$ 200 archives, up to 1,8M docs or 5.2M pages (700K digitalized), among text (handwritten and printed), letters, memos, diaries, images and videos.
- **Oral History Program (PHO)** A huge set of testimonies (in audio and video) consisting of more than 2K interviews, which correspond to up to 6K hours of recordings. 90% in digital format. Only 10% is transcribed. Limit access, not online.
- **Brazilian Historical Biographic Dictionary (DHBB)** 7,5K entries, 6,5K are of biographical and 1K related to institutions, events and concepts of interest for the Brazilian history after 1930. Carefully revised entries by researchers. Few metadata.

Currently Architecture



Currently Relational DB

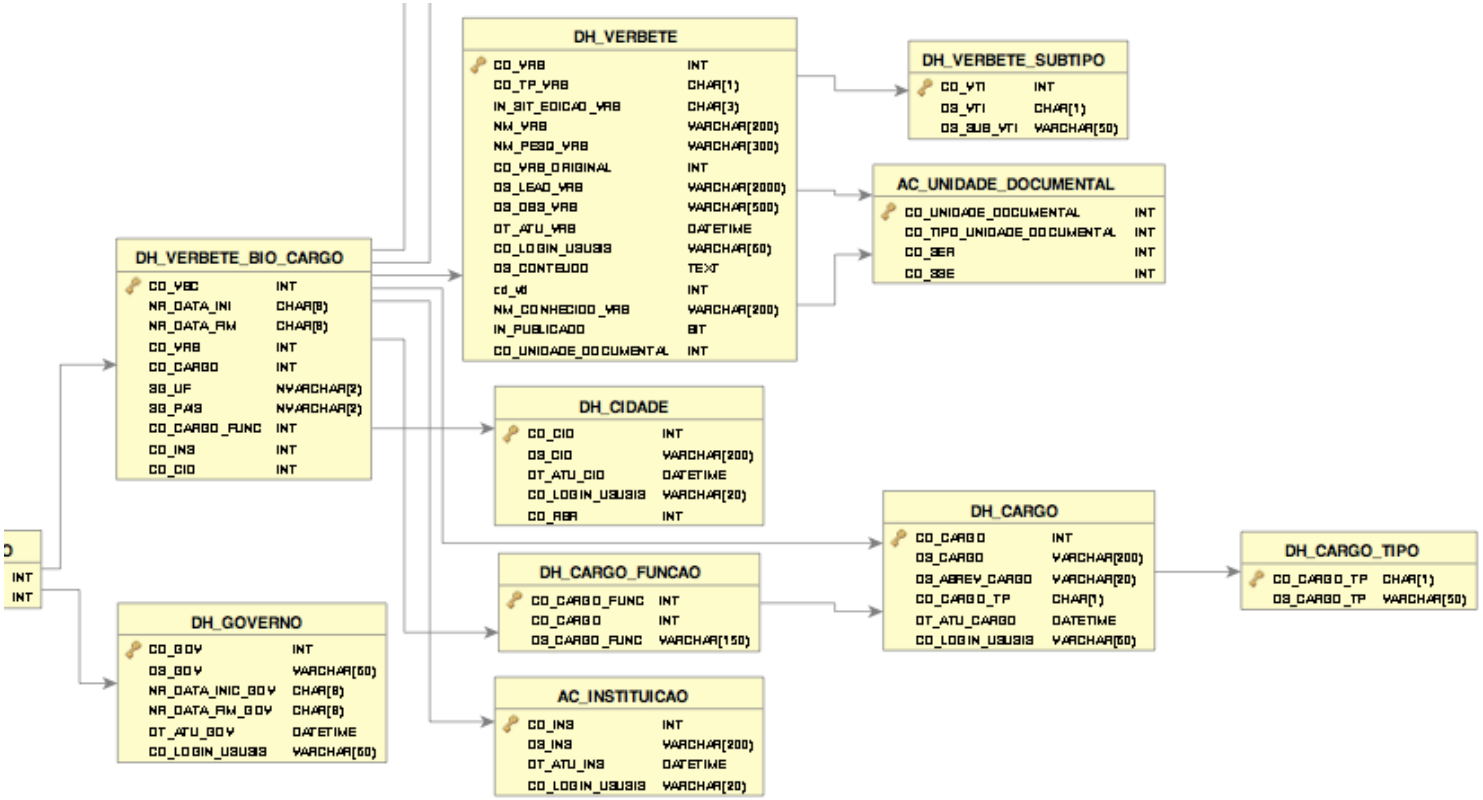
89 tables/classes and 660 columns/properties.



Problems

- Currently architecture is hard and costly to maintain and improve given the relational model nature and systems;
- innovative initiatives are usually postponed;
- The data is available online but on the "deep web";
- CPDOC's do not adopt any standard data model or vocab: (1) inhibit interoperability with other open resources; and (2) hardly useful for people outside CPDOC.
- data files (audio, videos and images) scattered in different file servers, DB only stores metadata and file paths (loose coupling).

Some inconsistencies



"verbete" is a dictionary entry. "bio_cargo" is a position ("cargo") that the described person had during a specific time during which he/she carried on a particular assignment ("funcao"). Controled lists but no standards! Double relation between "bio_cargo" and "cargo".

Inconsistencies are not always straightforward to fix

```
DELETE {
 ?bioc cpdoc:dbo_DH_VERBETE_BIO_CARGO_CD_CARGO ?cargo
}
INSERT {
  graph <http://cpdoc.fgv.br/sys/update1/> {
    ?bioc cpdoc:dbo_DH_VERBETE_BIO_CARGO_CD_CARGO_FUNC _:funcao .
   _:funcao rdf:type  cpdoc:dbo_DH_CARGO_FUNCAO ;
            cpdoc:dbo_DH_CARGO_FUNCAO_CD_CARGO ?cargo .
  }
}
WHERE {
 ?bioc cpdoc:dbo_DH_VERBETE_BIO_CARGO_CD_CARGO ?cargo .
 filter not exists {
  ?bioc cpdoc:dbo_DH_VERBETE_BIO_CARGO_CD_CARGO_FUNC ?cf .
  ?cf   cpdoc:dbo_DH_CARGO_FUNCAO_CD_CARGO ?cargo .
 }
}
```

. . . when we recognize the battle against chaos, mess, and unmastered complexity as one of computing science's major callings, we must admit that "Beauty is our Business". (Edsger W. Dijkstra)

Some beautiful arguments using mathematical induction. `http://goo.gl/KQ9j7Q`.

The Long Run Project

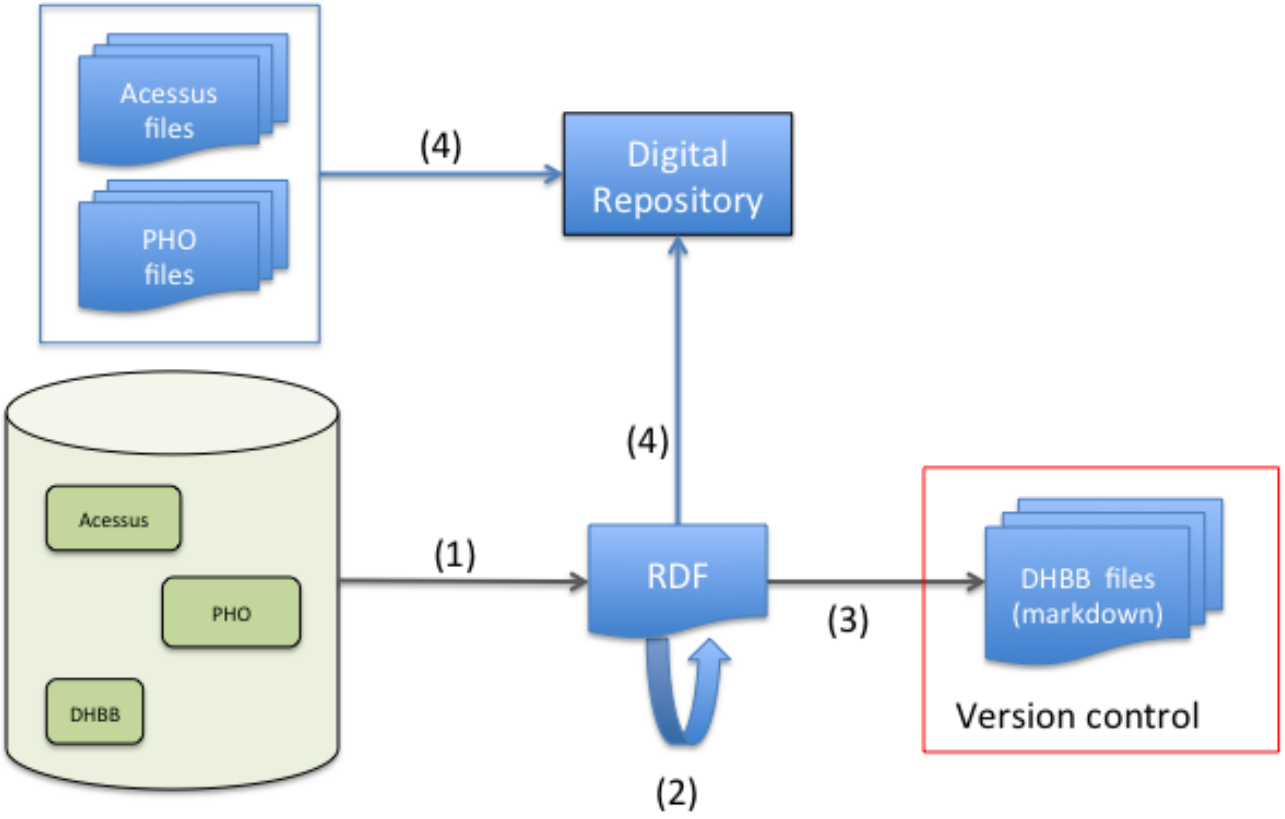
- Joint project between CPDOC and EMAp (Mathematical School);
- Enrich the structure (semantics) of CPDOC data;
- Open and expose CPDOC's data and architecture making it more maintainable and dynamic;
- Uniform and integrated data treatment (standards and interlinks between collections).

Motivations

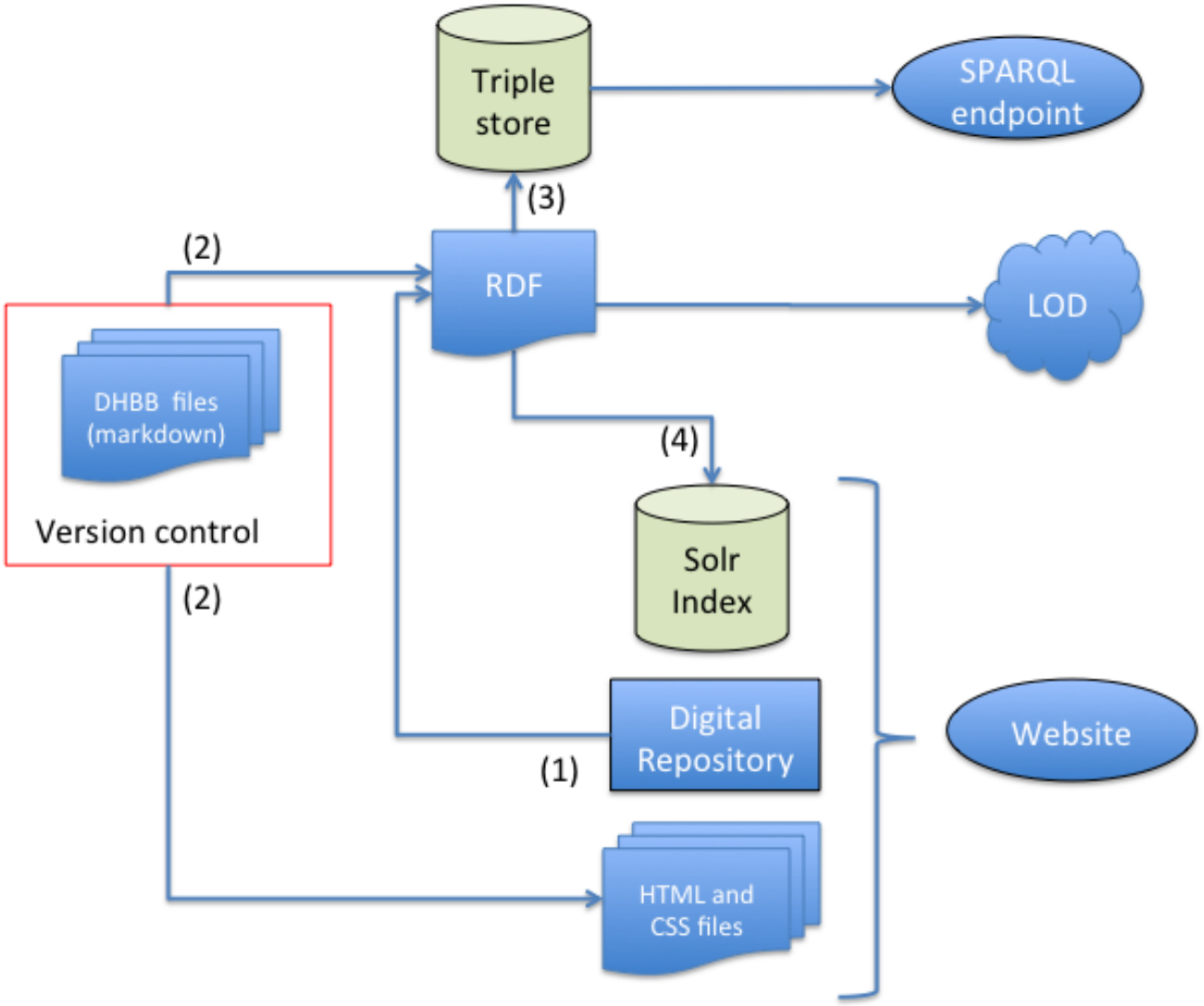
- Open Linked Data Initiative Principals;
- Distributed open source development model/tools (collaborative data maintenance and creation);
- From data owner to data curator;

The migration process

(1) D2RQ was extracted RDF from relational; (2) enrichment of data semantics (next slides); (3) DHBB entries to simple markdown files with YAML headers; (4) PHO and Accessus collections are moved to DRMS (standards vocab, access control, faced search, long-term preservation, OAI-PMH support etc.

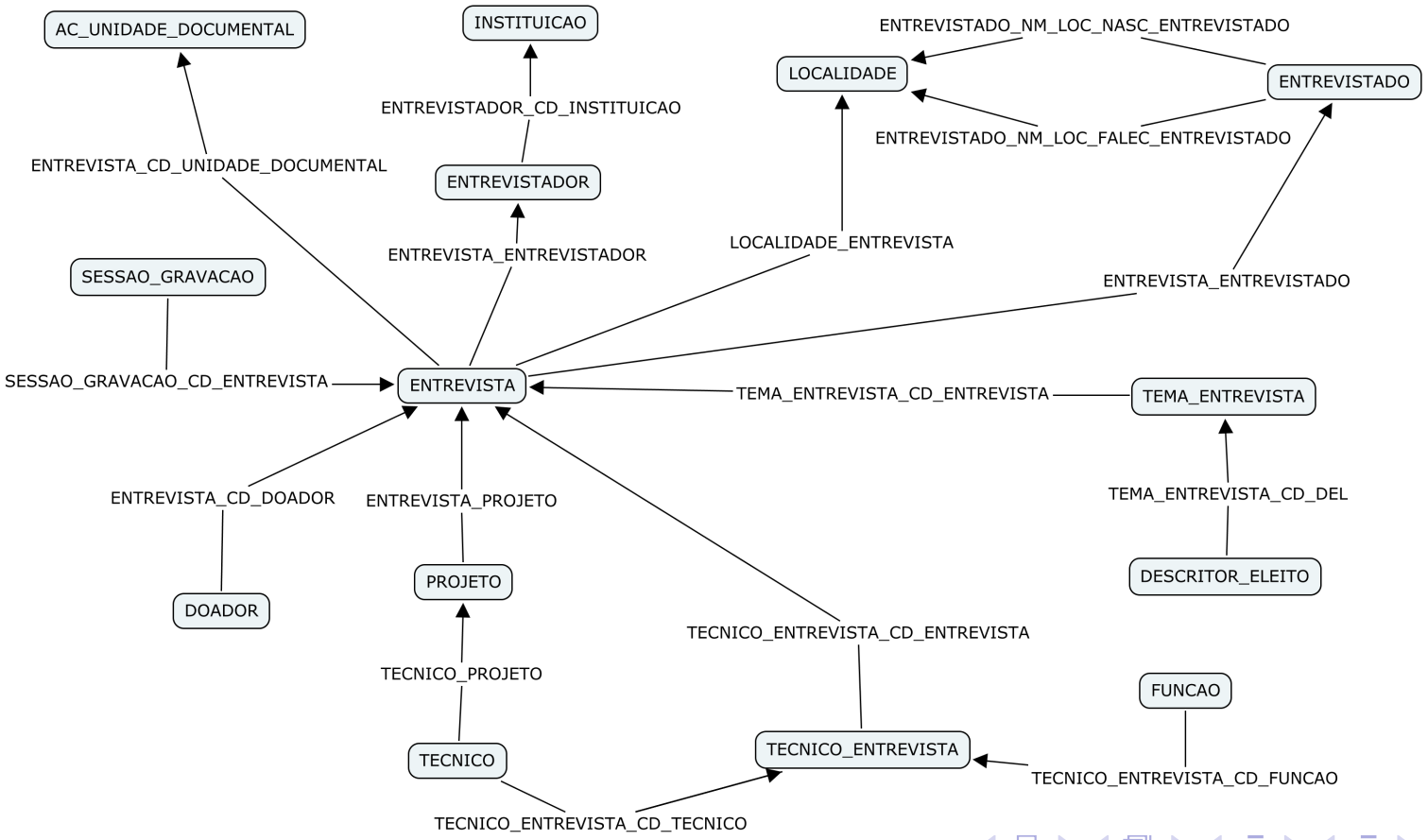


The desired architecture



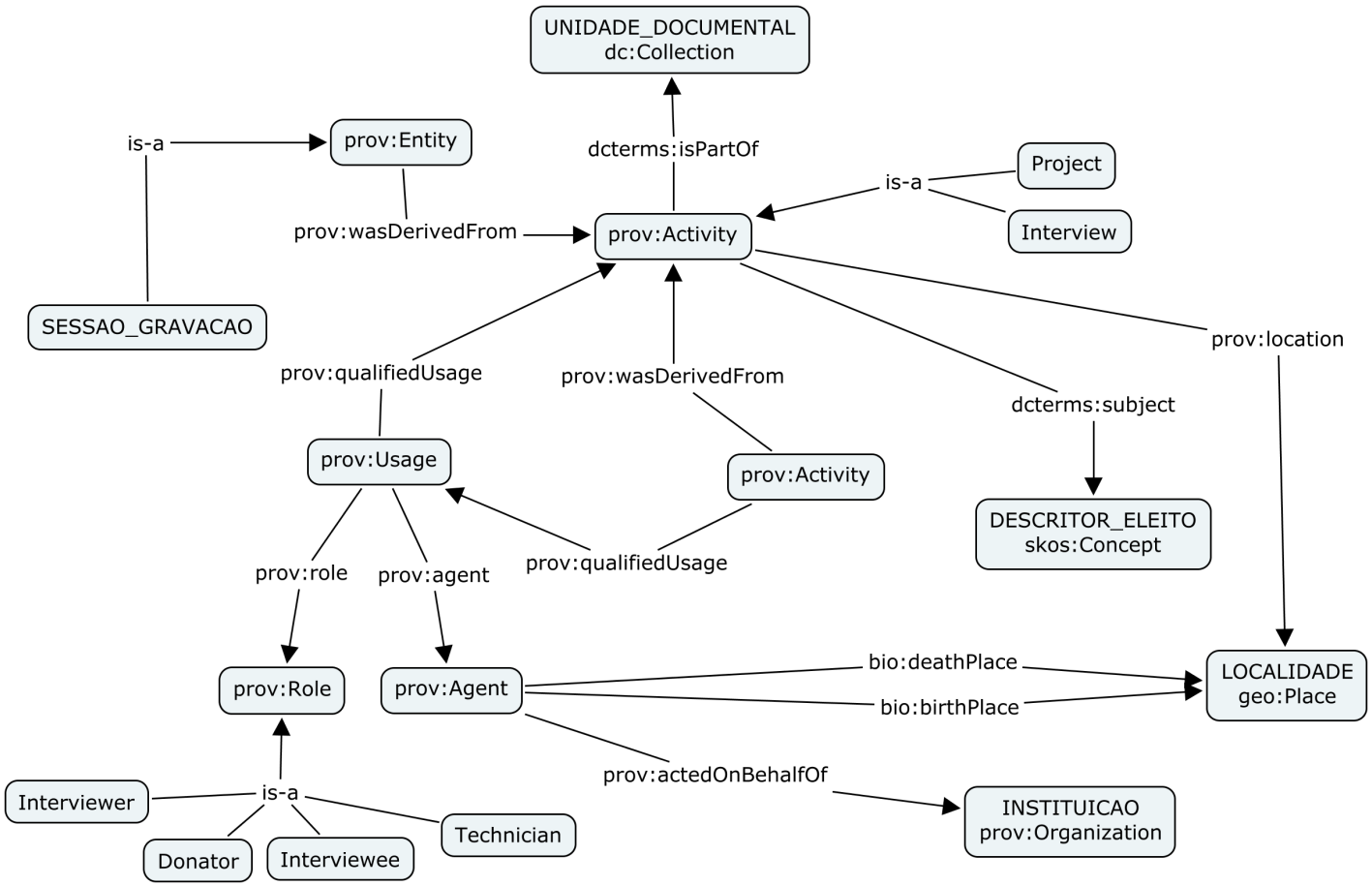
Improving semantics

1-1 with original relational DB. The connection of technician and interview is parameterized by different roles, the donator, interviewer and interviewed of an interview are modeled each one in a specific table. In this case interviewed, interviewer, donator and technician are all people ("ad hoc" modeling).



Improving semantics

prov centric but uses **skos**, **dc**, **foaf**, **bio** and **geo**, **frbr** etc. some classes can be subclasses of standard classes, `Interview`, some classes can be replaced by standard classes, `localidade`.



Conclusions

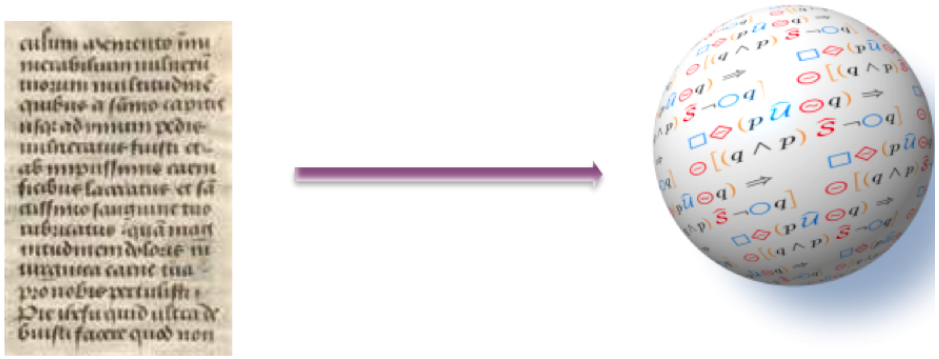
- Challenge 1: convince CPDOC researchers to make the transition to data owners to curators.
- Challenge 2: adapt researchers to new technologies (VC, text editors, scripts?, distributed workflow etc)
- Model refinements (corrections, transformations by alignments) can be not straightforward.
- Still a lot to be done. For instance...

Other Research Opportunities

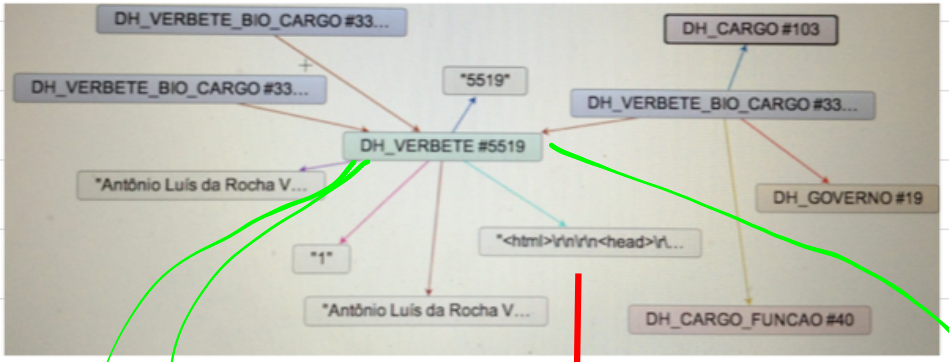
- Natural language processing: processing the DHBB entries to discover relations between entries and with other linked data and resources. DHBB for NLP and vice versa!
- Ontology alignmnent algorithms for (semi-)automated the model transformations.

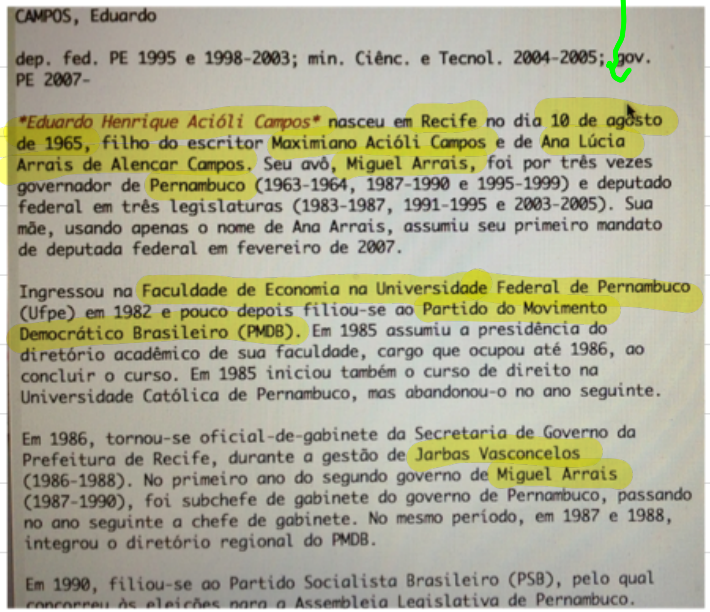
Natural Language Processing

- Manually discovered $\approx$ 50 links to dbpedia (Presidents of Brazil, presidents of the Senate, political parties etc.)
- NLP and text mining of DHBB entries: (1) proper names; (2) word sense disambiguation using the openWordnet-PT (lexical resource); and (3) named entity recognition and creation of links between DHBB entries.
- 133,036 proper names identified (some few mistakes). Potencially entities (people, locations, organizations etc)
- Use grammars, lexical resources, formal ontologies, and logical tools to reason about knowledge obtained from processing text in Portuguese (Computational Semantics: KB, KR, and ATP);



Natural Language Processing





CAMPOS, Eduardo

dep. fed. PE 1995 e 1998-2003; min. Ciênc. e Tecnol. 2004-2005; gov. PE 2007-

Eduardo Henrique Acióli Campos nasceu em Recife no dia 10 de agosto de 1965, filho do escritor Maximiano Acióli Campos e de Ana Lúcia Arrais de Alencar Campos. Seu avô, Miguel Arrais, foi por três vezes governador de Pernambuco (1963-1964, 1987-1990 e 1995-1999) e deputado federal em três legislaturas (1983-1987, 1991-1995 e 2003-2005). Sua mãe, usando apenas o nome de Ana Arrais, assumiu seu primeiro mandato de deputada federal em fevereiro de 2007.

Ingressou na Faculdade de Economia na Universidade Federal de Pernambuco (Ufpe) em 1982 e pouco depois filiou-se ao Partido do Movimento Democrático Brasileiro (PMDB). Em 1985 assumiu a presidência do diretório acadêmico de sua faculdade, cargo que ocupou até 1986, ao concluir o curso. Em 1985 iniciou também o curso de direito na Universidade Católica de Pernambuco, mas abandonou-o no ano seguinte.

Em 1986, tornou-se oficial-de-gabinete da Secretaria de Governo da Prefeitura de Recife, durante a gestão de Jarbas Vasconcelos (1986-1988). No primeiro ano do segundo governo de Miguel Arrais (1987-1990), foi subchefe de gabinete do governo de Pernambuco, passando no ano seguinte a chefe de gabinete. No mesmo período, em 1987 e 1988, integrou o diretório regional do PMDB.

Em 1990, filiou-se ao Partido Socialista Brasileiro (PSB), pelo qual concorreu às eleições para a Assembleia Legislativa de Pernambuco.

Audio and Transcriptions

Sinal processing to (semi-) automatic produce transcriptions, alignment with already available transcriptions and audio segmentation (interviewer/inverviwed);

Entrevista: 13.07.2004

(00:00:13,6) L.H – Bom, então vamos começar. O seu currículo é bastante sucinto, são dados gerais...

(00:00:21,2) A.P – Ah, mas pode cortar.

(00:00:23,1) L.H. – Não, a gente não quer cortar, a gente quer acrescentar. A gente queria começar do começo: quando o senhor nasceu, onde, como era a família…

(00:00:34,7) A.P. – Ah, vem de lá?

(00:00:35,8) L.H – Vem de lá, *from the begining.*

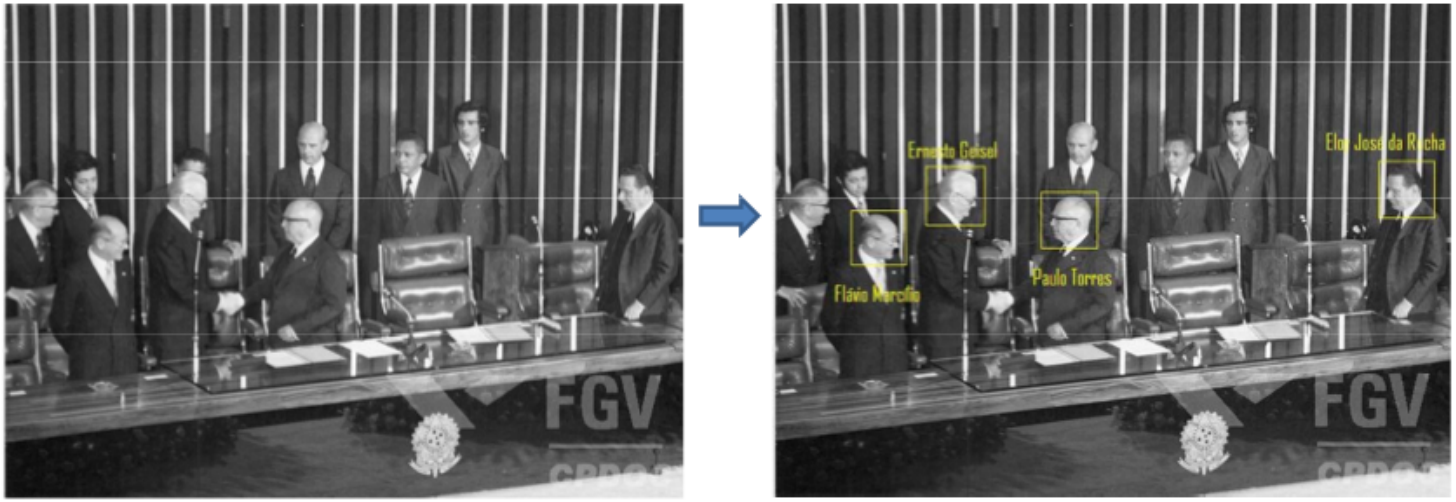
(00:00:39,2) A.P – Família de imigrantes, todos nós acho que somos.

(00:00:45,5) L.H. – Imigrantes italianos?

(00:00:46,4) A.P. – Meu pai saiu da Itália, tinha dois navios no inverno na Itália. A Itália é muito fria. Em Veneza, onde ele morava, é mais frio ainda por causa da umidade. E os dois navios que vinham para a América, um era preto e o outro era cinza. O navio preto tinha uma fila

Faces recognition and identification

Image processing techniques to face recognition in photos collections.



Legend: Esq./dir.: (1o plano) Flávio Marcílio (1o); Ernesto Geisel (2o); Paulo Torres (3o); Eloy José da Rocha (4o). (2o plano) Adalberto Pereira dos Santos (1o). Foto: Agência Nacional (Estúdio/Agência).

Obrigado!

S: (v) thank, give thanks (express gratitude or show appreciation to)

```
(=>
  (and
      (instance ?THANK Thanking)
      (agent ?THANK ?AGENT)
      (patient ?THANK ?THING)
      (destination ?THANK ?PERSON))
  (and
      (instance ?PERSON Human)
      (or
          (holdsDuring
              (WhenFn ?THANK)
              (wants ?AGENT ?THING))
          (holdsDuring
              (WhenFn ?THANK)
              (desires ?AGENT ?THING)))))
```

SUMO Ontology, http://www.ontologyportal.org