



Semantic Lattes and VIVO

Alexandre Rademaker
IBM Research and FGV/EMAp
and
Edward Hermann Haeusler
PUC-Rio

Introduction

- PhD 2010
 - Computer Science
 - Proof Theory, Description Logics, ATP
 - Knowledge Representation and Reasoning
 - Ontologies Alignments, Instance Matching etc.
- FGV 1996-2010.
 - 1996-2010 IT/Supporting Researchers
 - 2010-? Professor/Researcher at EMAp
- IBM Research Brazil: started Dec 2012

IBM Research Brazil



Brazil Lab was created in 2010.



12 labs. 6 continents.



Getulio Vargas Foundation

School of Applied Mathematics

<http://emap.fgv.br>

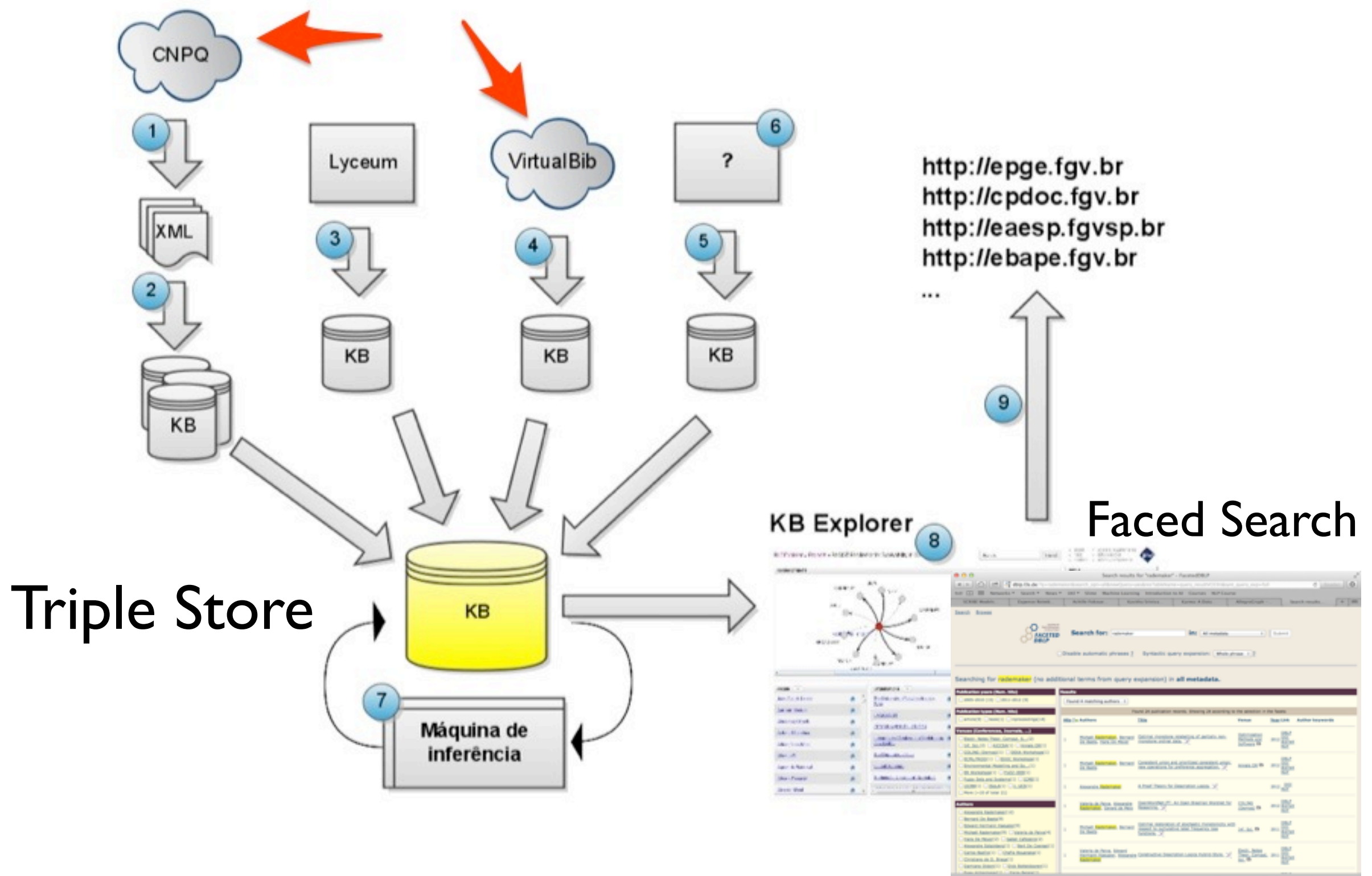
“**Fundação Getulio Vargas (FGV)** is a Brazilian higher education and research institution founded in December 20, 1944. It offers regular courses of Economics, Business Administration, Law, Social Sciences and **Applied Mathematics**. Its original goal was to train people for the country's public-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



The Project

- Almost all FGV departments have to deal with publications and researchers profiles in their websites. Duplication of Efforts!
- The FGV's administration need a “big picture” of the research activities and in-house skills.
- All FGV departments have to provide the same reports: for FGV's administration, CAPES (Government agency that rank pos-graduate courses and departments across the country etc)
- Started in mid of 2009!

Lattes@FGV architecture



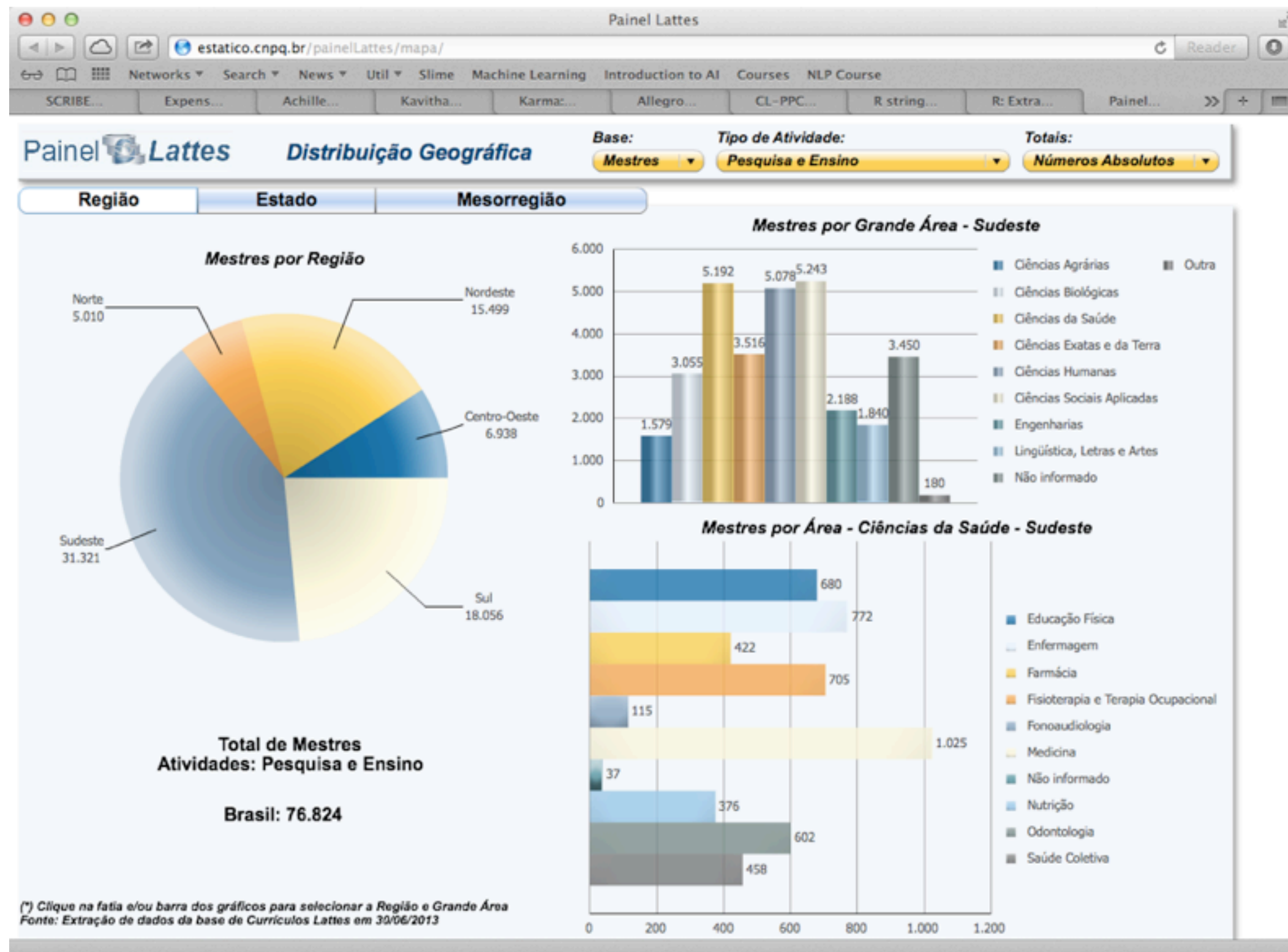
Lattes Platform

- Brazilian Government initiative
- <http://lattes.cnpq.br>

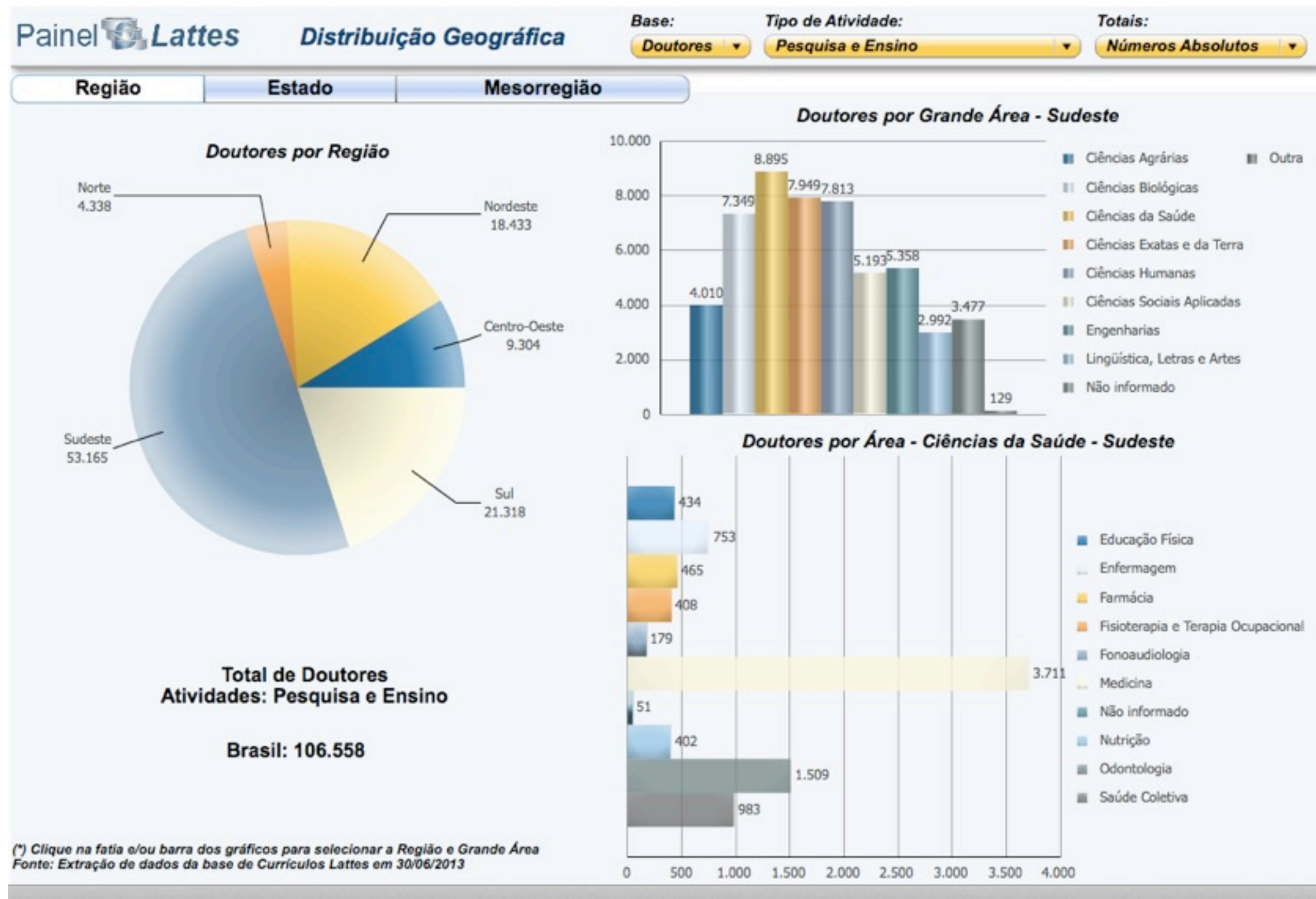
The Lattes Platform is a online system used by almost all researchers in Brazil to maintain their curriculum vitae. Developed by CNPq (National Council for Scientific and Technological Development) in the mid-80s, the platform is an instrument for guide investments in research in Brazil and evaluate the brazilian research community.

Having an updated Lattes Resume is eligibility precondition for proposal submissions for public investment.

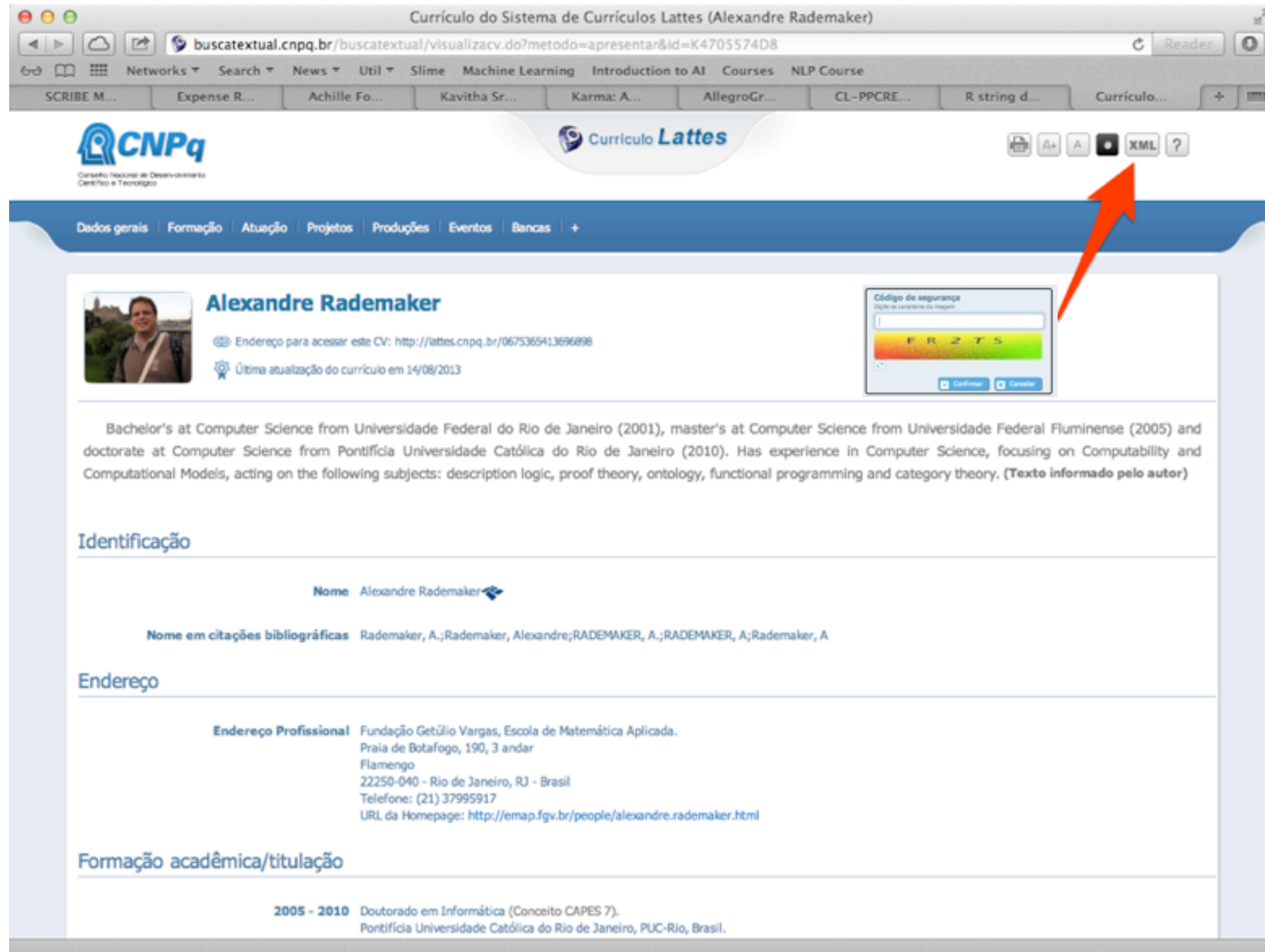
Lattes Platform



Lattes Platform



Lattes Platform



The screenshot shows a web browser window displaying the Lattes Platform profile for Alexandre Rademaker. The browser's address bar shows the URL: `buscatextual.cnpq.br/buscatextual/visualizacv.do?metodo=apresentar&id=K4705574D8`. The browser's tabs include "SCRIBE M...", "Expense R...", "Achille Fo...", "Kavitha Sr...", "Karma: A...", "AllegroGr...", "CL-PPCRE...", "R string d...", and "Currículo...". The Lattes Platform header features the CNPq logo, the "Currículo Lattes" title, and a navigation menu with links: "Dados gerais", "Formação", "Atuação", "Projetos", "Produções", "Eventos", "Bancas", and a "+" icon. A red arrow points to the "XML" button in the top right corner of the profile page. The profile itself includes a photo of Alexandre Rademaker, his name, a link to his CV, and the date of the last update (14/08/2013). A security code verification box is also visible. The profile text describes his education (Bachelor's, Master's, and Doctorate in Computer Science) and his research interests (Computability and Computational Models). The "Identificação" section lists his name and bibliographic citations. The "Endereço" section provides his professional address, phone number, and homepage URL. The "Formação acadêmica/titulação" section lists his doctorate from PUC-Rio in 2010.

Currículo do Sistema de Currículos Lattes (Alexandre Rademaker)

buscatextual.cnpq.br/buscatextual/visualizacv.do?metodo=apresentar&id=K4705574D8

Networks Search News Util Slime Machine Learning Introduction to AI Courses NLP Course

SCRIBE M... Expense R... Achille Fo... Kavitha Sr... Karma: A... AllegroGr... CL-PPCRE... R string d... Currículo...

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

Currículo Lattes

Dados gerais | Formação | Atuação | Projetos | Produções | Eventos | Bancas | +

Alexandre Rademaker

Endereço para acessar este CV: <http://lattes.cnpq.br/0675365413696898>

Última atualização do currículo em 14/08/2013

Bachelor's at Computer Science from Universidade Federal do Rio de Janeiro (2001), master's at Computer Science from Universidade Federal Fluminense (2005) and doctorate at Computer Science from Pontifícia Universidade Católica do Rio de Janeiro (2010). Has experience in Computer Science, focusing on Computability and Computational Models, acting on the following subjects: description logic, proof theory, ontology, functional programming and category theory. (Texto informado pelo autor)

Identificação

Nome: Alexandre Rademaker

Nome em citações bibliográficas: Rademaker, A.; Rademaker, Alexandre; RADEMAKER, A.; RADEMAKER, A; Rademaker, A

Endereço

Endereço Profissional Fundação Getúlio Vargas, Escola de Matemática Aplicada.
Praia de Botafogo, 190, 3 andar
Flamengo
22250-040 - Rio de Janeiro, RJ - Brasil
Telefone: (21) 37995917
URL da Homepage: <http://emap.fgv.br/people/alexandre.rademaker.html>

Formação acadêmica/titulação

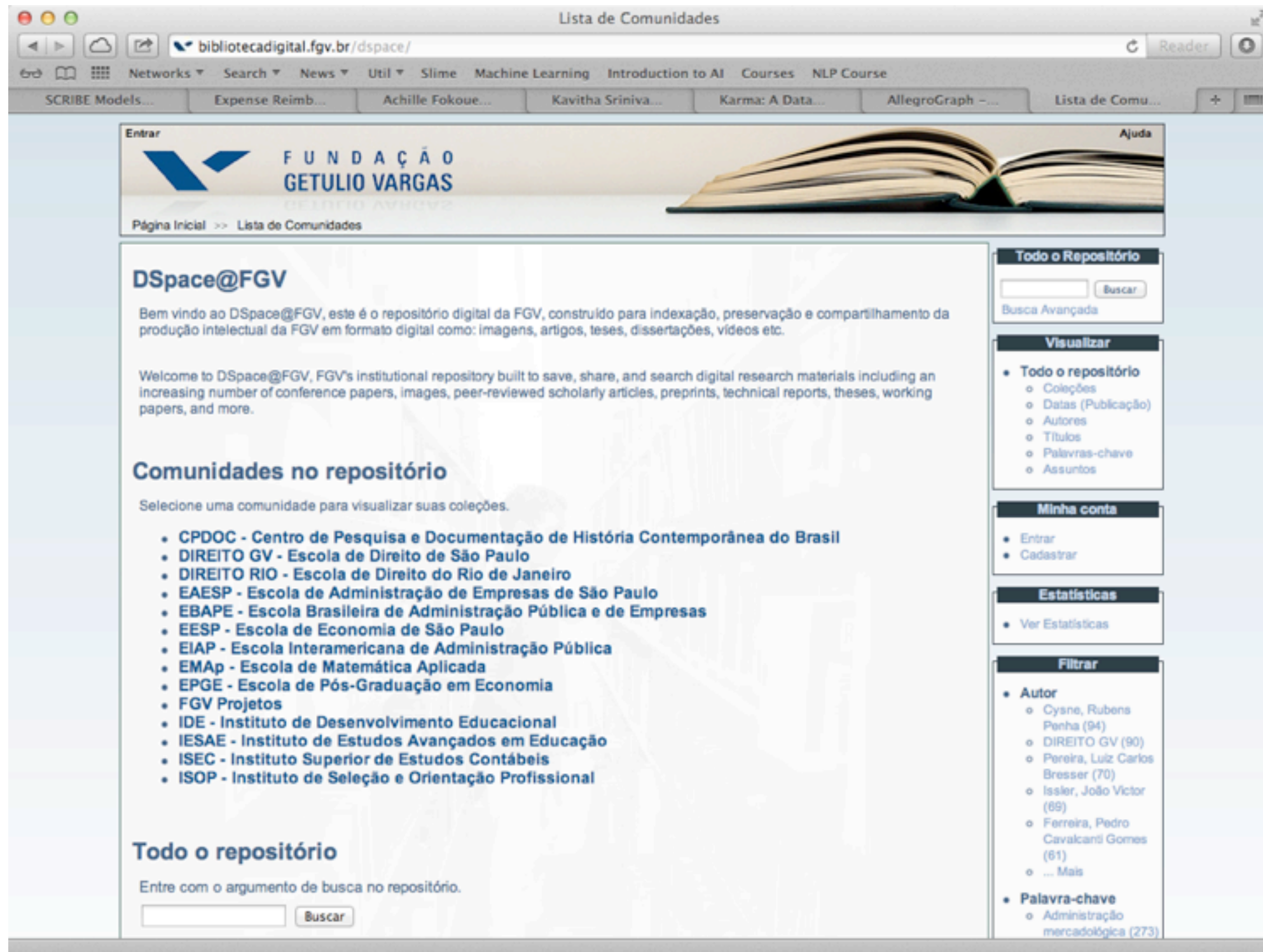
2005 - 2010 Doutorado em Informática (Conceito CAPES 7).
Pontifícia Universidade Católica do Rio de Janeiro, PUC-Rio, Brasil.

Lattes good and bad

- Good source of information that research must keep updated!
- It doesn't adopted (semantic) standards besides data formats (XML)
- Data is not really in open-access model! We can parse HTML from CNPq site or Institutions must sign an agreement for accessing CV from their researchers (XML).
- Started with a promise to be driven by the researchers community but ends up begin driven by the government.

<http://lattes.cnpq.br/lattes/> (not updated!)

FGV Digital Library



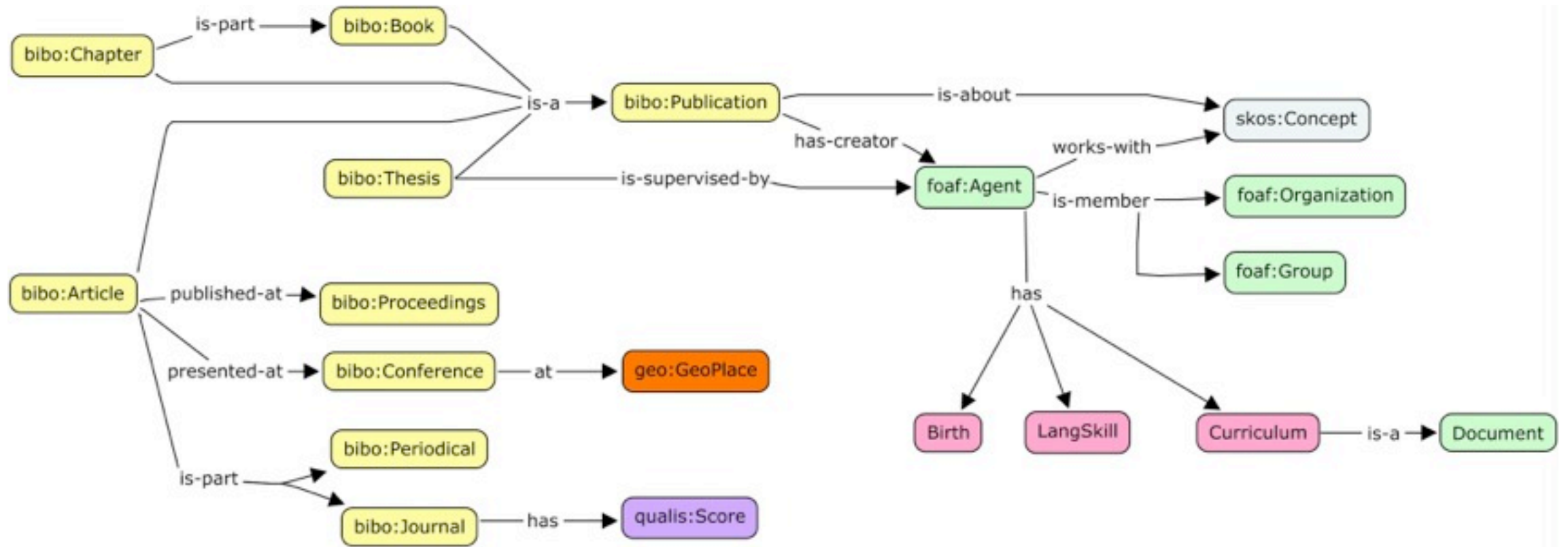
OAI-PMH Interface ... RDF

XML to RDF (xslt)

```
▼<CURRICULO-VITAE SISTEMA-ORIGEM-XML="LATTES_OFFLINE" DATA-ATUALIZACAO="09082011" HORA-ATUALIZACAO="125948" NUMERO-IDENTIFICADOR="9361065863868373">
▼<DADOS-GERAIS NOME-COMPLETO="Luiz Carlos Di Serio" NOME-EM-CITACOES-BIBLIOGRAFICAS="Di SERIO, Luiz Carlos" NACIONALIDADE="B" PAIS-DE-
NASCIMENTO="Brasil" UF-NASCIMENTO="SP" CIDADE-NASCIMENTO="DUARTINA" DATA-NASCIMENTO="12051949" SEXO="MASCULINO" NOME-DO-PAI="ALCYR PAULO DI SERIO"
NOME-DA-MAE="ALEXANDRA ZANATTA DI SERIO" PERMISSAO-DE-DIVULGACAO="NAO" DATA-FALECIMENTO="">
  <RESUMO-CV TEXTO-RESUMO-CV-RH="..." TEXTO-RESUMO-CV-RH-EN="..." />
  <OUTRAS-INFORMACOES-RELEVANTES OUTRAS-INFORMACOES-RELEVANTES="..." />
  <ENDEREÇO FLAG-DE-PREFERENCIA="ENDEREÇO_INSTITUCIONAL">...</ENDEREÇO>
  <FORMACAO-ACADEMICA-TITULACAO>...</FORMACAO-ACADEMICA-TITULACAO>
  <ATUACOES-PROFISSIONAIS>...</ATUACOES-PROFISSIONAIS>
  <AREAS-DE-ATUACAO>...</AREAS-DE-ATUACAO>
  <IDIOMAS>...</IDIOMAS>
  <PREMIOS-TITULOS>...</PREMIOS-TITULOS>
</DADOS-GERAIS>
▼<PRODUCAO-BIBLIOGRAFICA>
  <TRABALHOS-EM-EVENTOS>...</TRABALHOS-EM-EVENTOS>
  <ARTIGOS-PUBLICADOS>...</ARTIGOS-PUBLICADOS>
  <LIVROS-E-CAPITULOS>
    <LIVROS-PUBLICADOS-OU-ORGANIZADOS>...</LIVROS-PUBLICADOS-OU-ORGANIZADOS>
    <CAPITULOS-DE-LIVROS-PUBLICADOS>...</CAPITULOS-DE-LIVROS-PUBLICADOS>
  </LIVROS-E-CAPITULOS>
  <TEXTOS-EM-JORNAIS-OU-REVISTAS>...</TEXTOS-EM-JORNAIS-OU-REVISTAS>
  <DEMAIS-TIPOS-DE-PRODUCAO-BIBLIOGRAFICA>...</DEMAIS-TIPOS-DE-PRODUCAO-BIBLIOGRAFICA>
  <ARTIGOS-ACEITOS-PARA-PUBLICACAO>...</ARTIGOS-ACEITOS-PARA-PUBLICACAO>
</PRODUCAO-BIBLIOGRAFICA>
  <PRODUCAO-TECNICA>...</PRODUCAO-TECNICA>
▼<OUTRA-PRODUCAO>
  <ORIENTACOES-CONCLUIDAS>...</ORIENTACOES-CONCLUIDAS>
  <DEMAIS-TRABALHOS SEQUENCIA-PRODUCAO="69">...</DEMAIS-TRABALHOS>
  <DEMAIS-TRABALHOS SEQUENCIA-PRODUCAO="239">...</DEMAIS-TRABALHOS>
</OUTRA-PRODUCAO>
▼<DADOS-COMPLEMENTARES>
  <FORMACAO-COMPLEMENTAR>...</FORMACAO-COMPLEMENTAR>
  <PARTICIPACAO-EM-BANCA-TRABALHOS-CONCLUSAO>...</PARTICIPACAO-EM-BANCA-TRABALHOS-CONCLUSAO>
  <PARTICIPACAO-EM-EVENTOS-CONGRESSOS>...</PARTICIPACAO-EM-EVENTOS-CONGRESSOS>
  <ORIENTACOES-EM-ANDAMENTO>...</ORIENTACOES-EM-ANDAMENTO>
  <INFORMACOES-ADICIONAIS-INSTITUICOES>...</INFORMACOES-ADICIONAIS-INSTITUICOES>
  <INFORMACOES-ADICIONAIS-CURSOS>...</INFORMACOES-ADICIONAIS-CURSOS>
</DADOS-COMPLEMENTARES>
</CURRICULO-VITAE>
```

<https://github.com/arademaker/slattes/>

Target Model

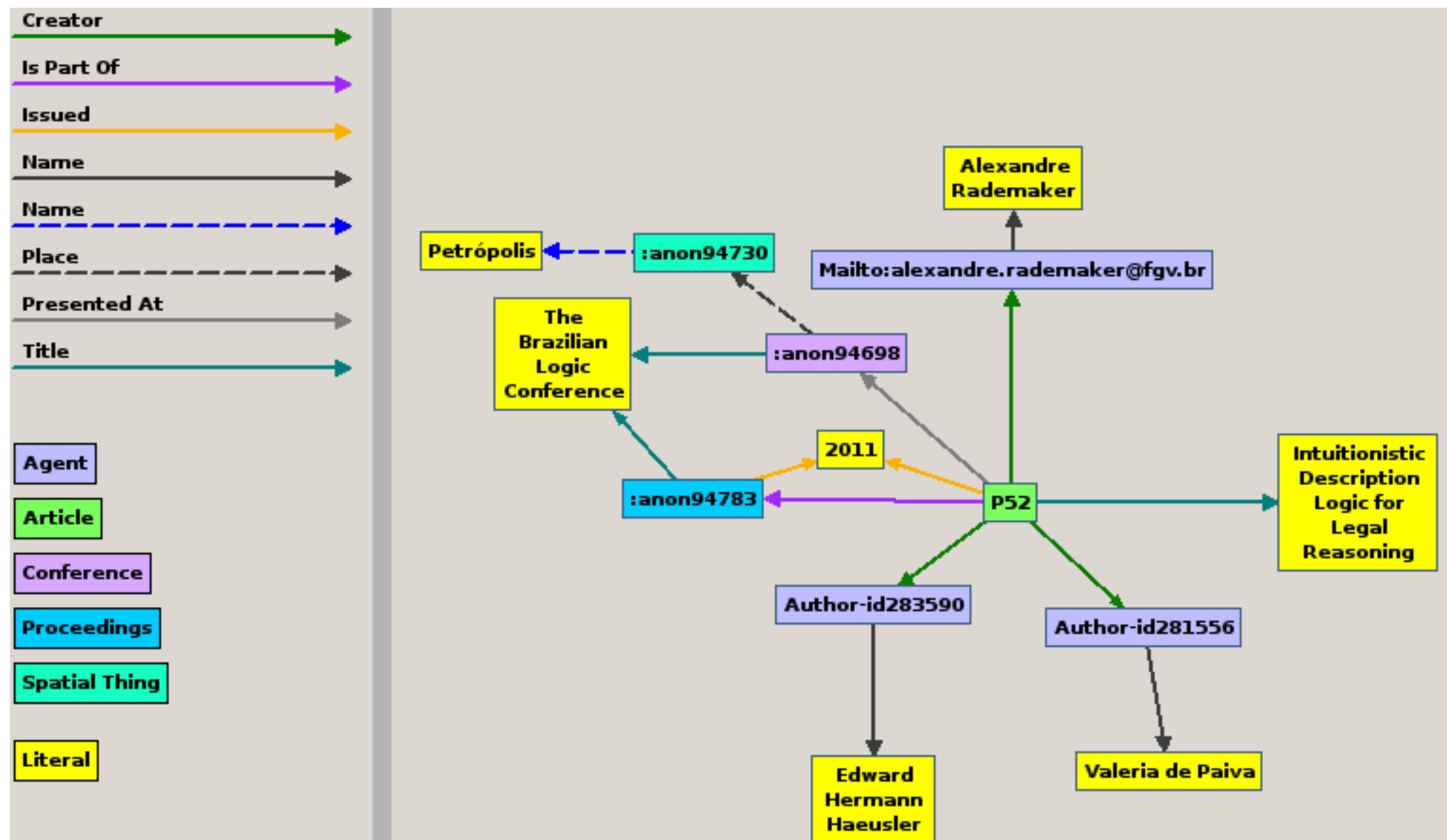


Vocabularies and Ontologies: foaf, dc, bibo, geo, skos, bio etc

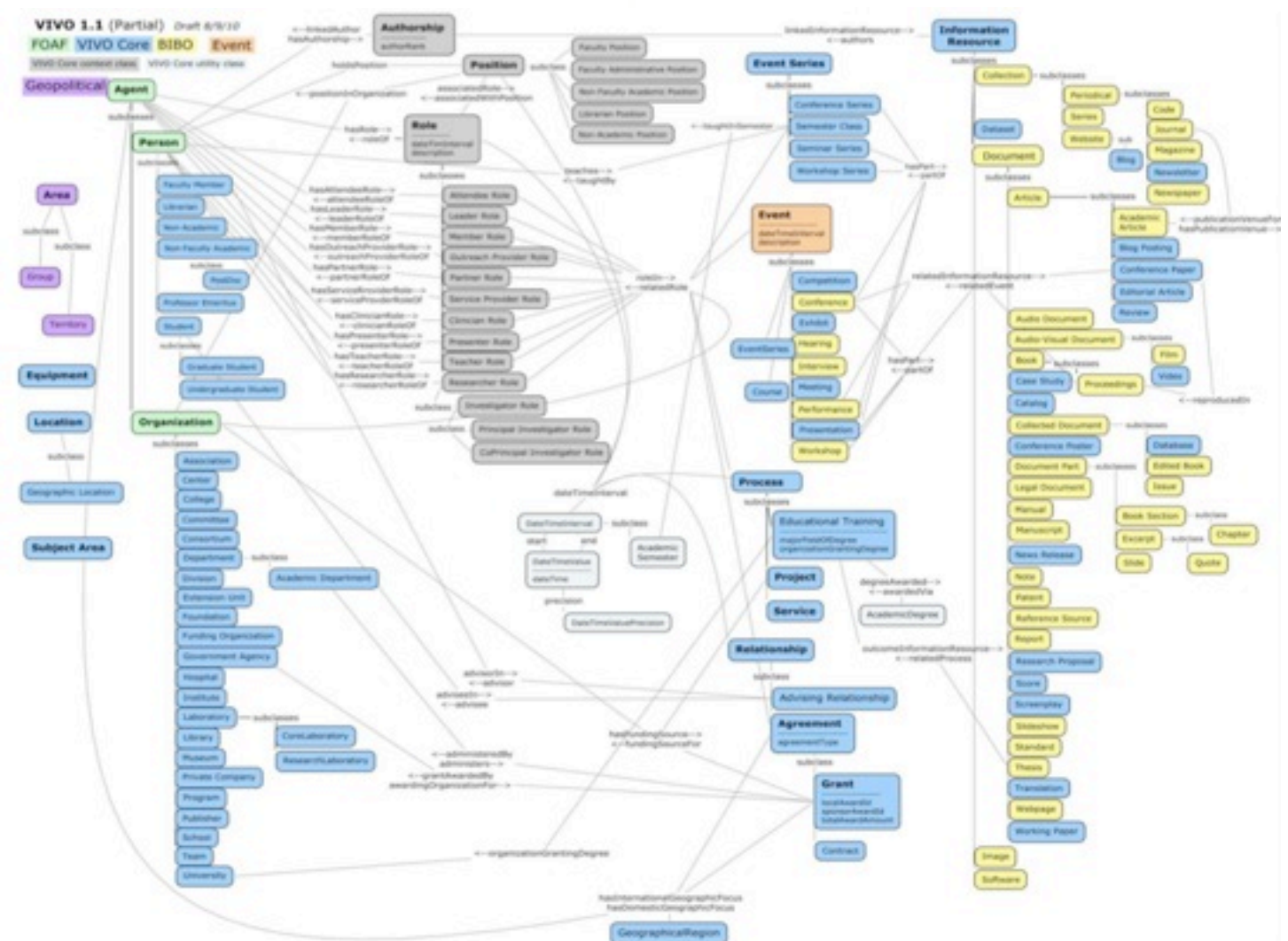
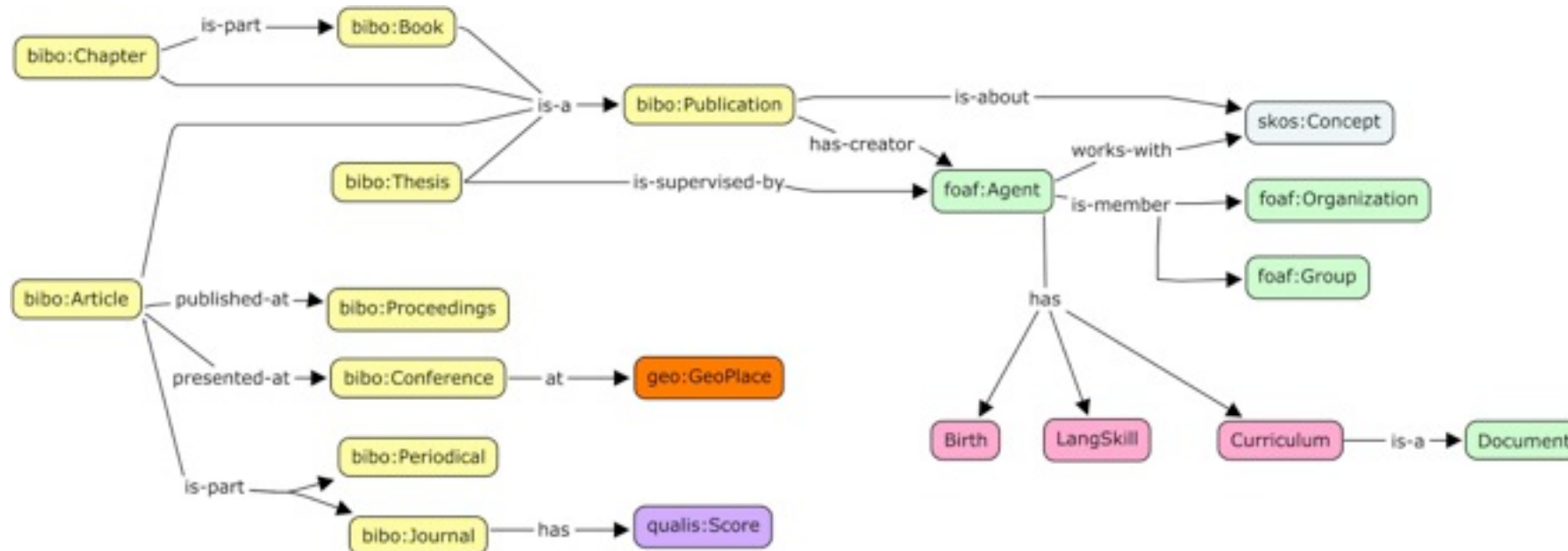
Graph fragment

Repository lattes — 1,793,017 statements

Sparql Endpoint: <http://logics.emap.fgv.br:10035/repositories/lattes>



VIVO Alignment?



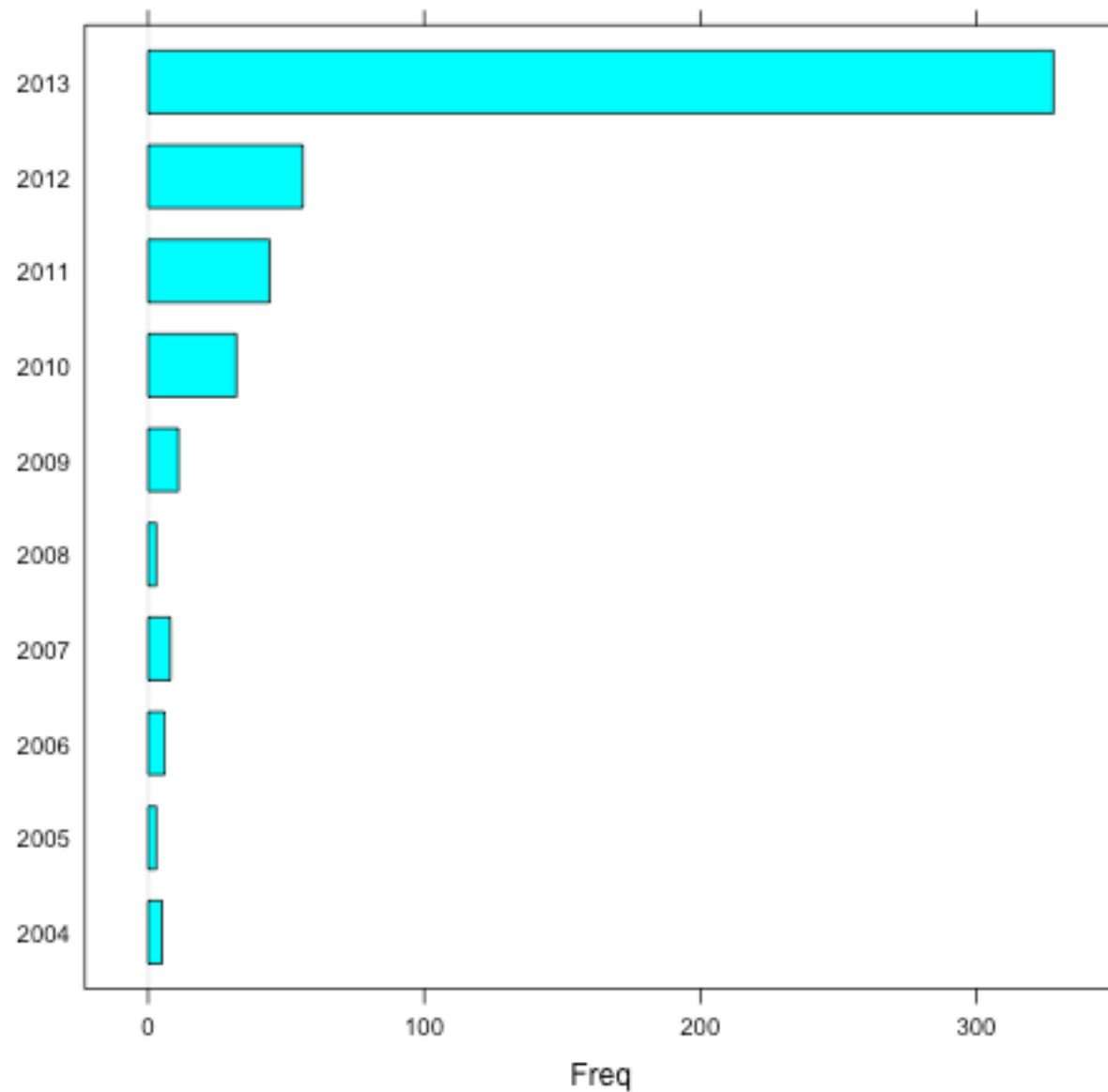
Not far from being easily used by:

<http://research.icts.uiowa.edu/polyglot/>

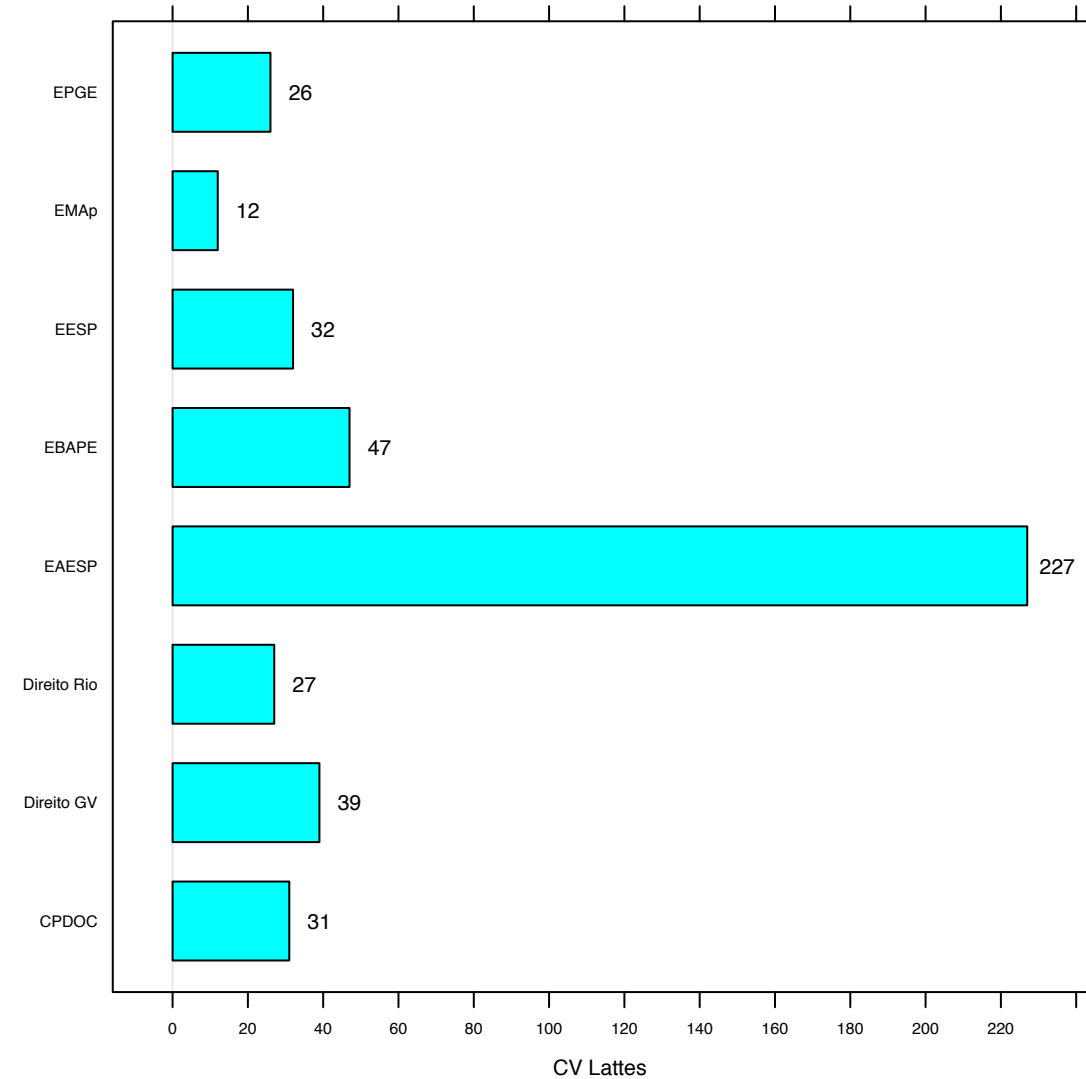
<http://beta.vivosearch.org>

Some reports

CV / last update

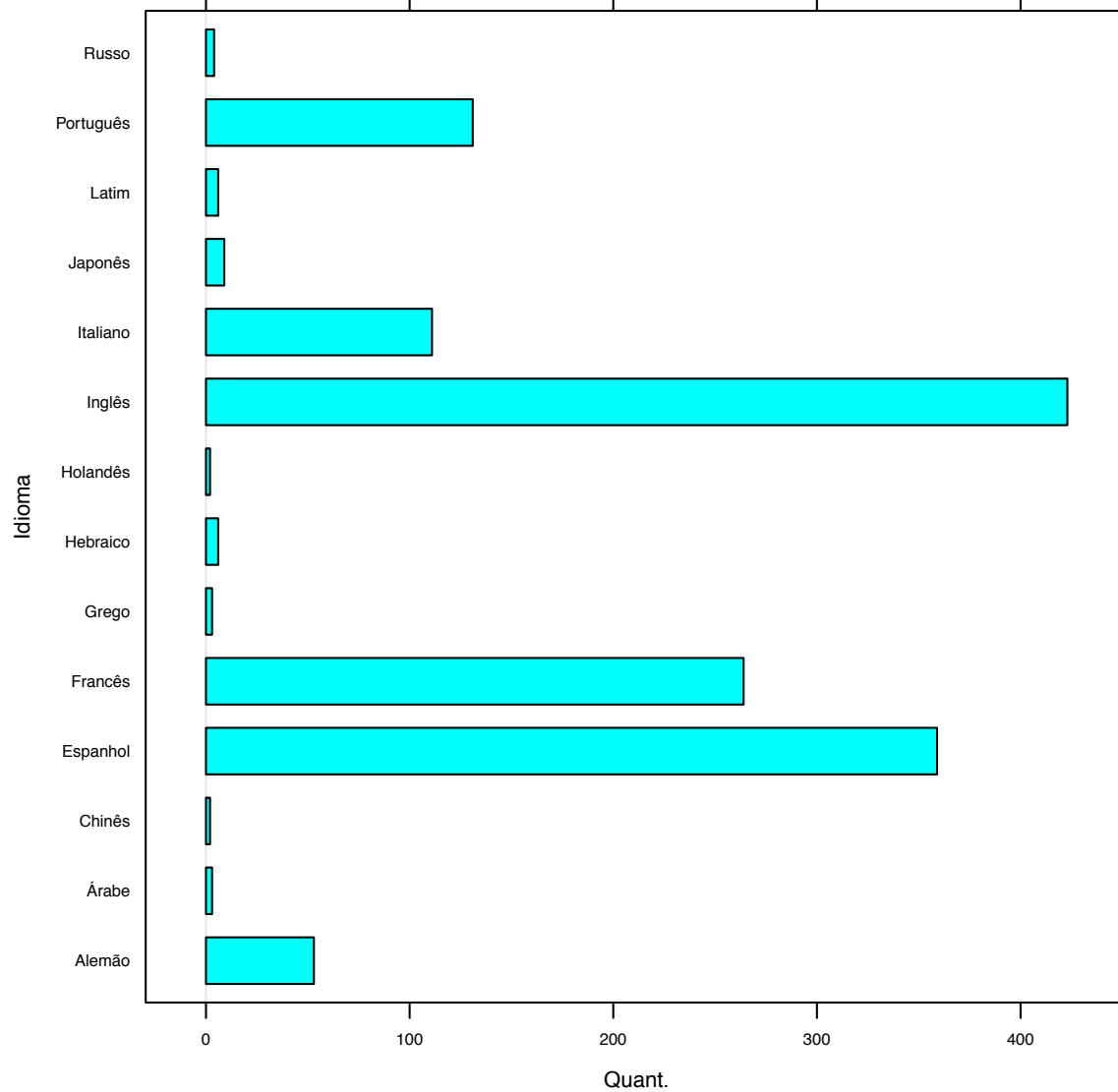


CVs per Department

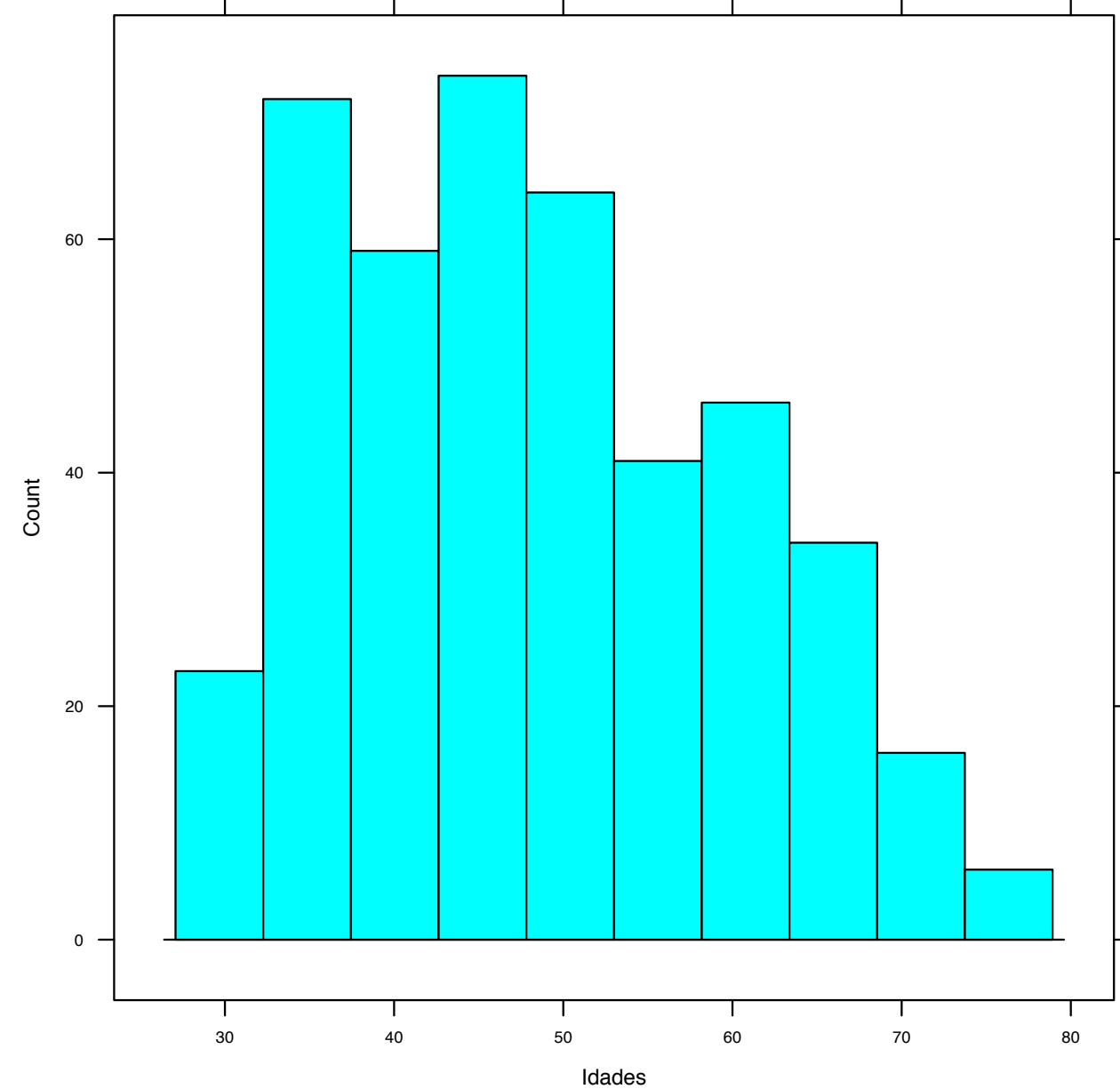


More reports

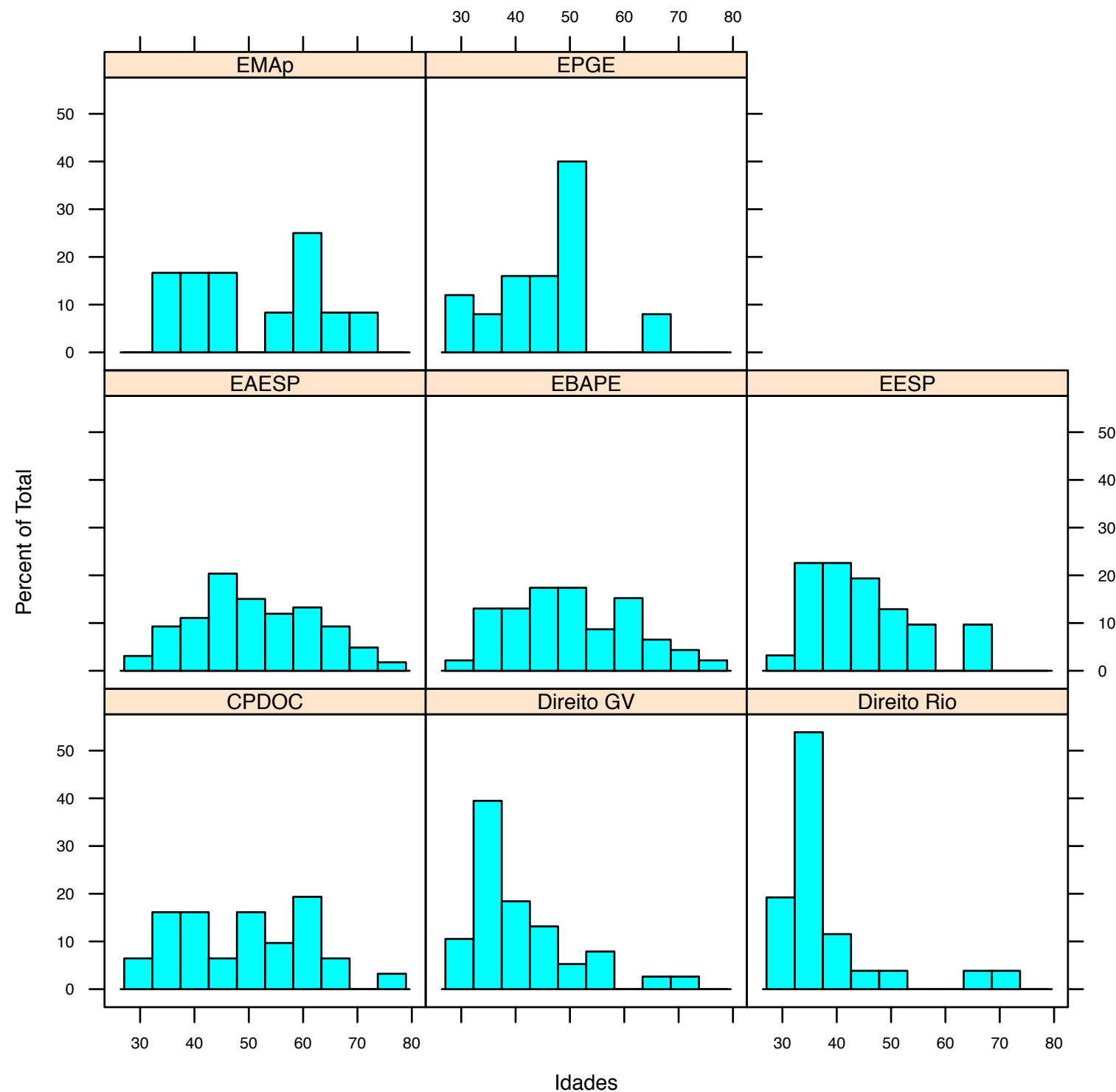
Language skills



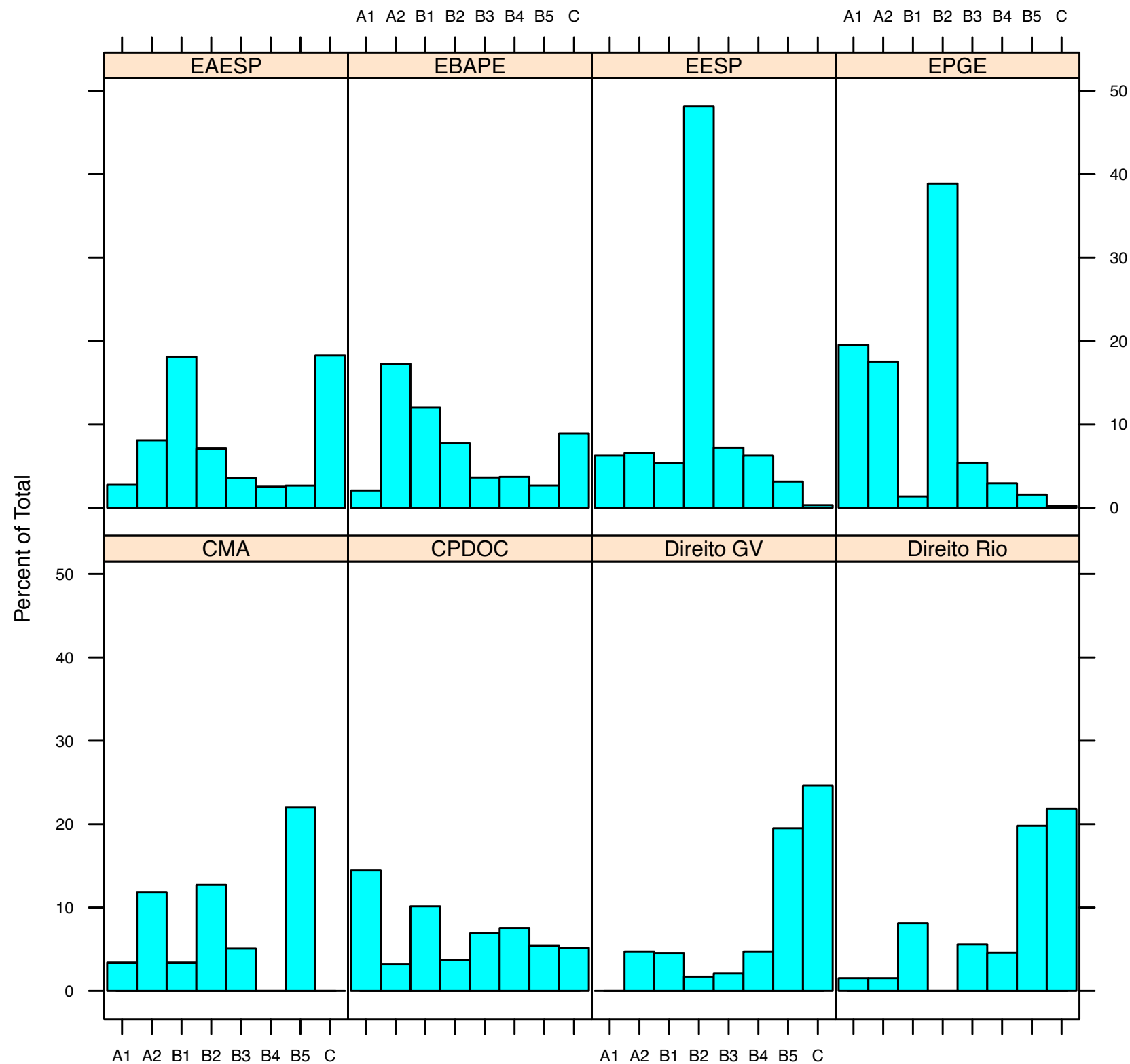
How old are we?



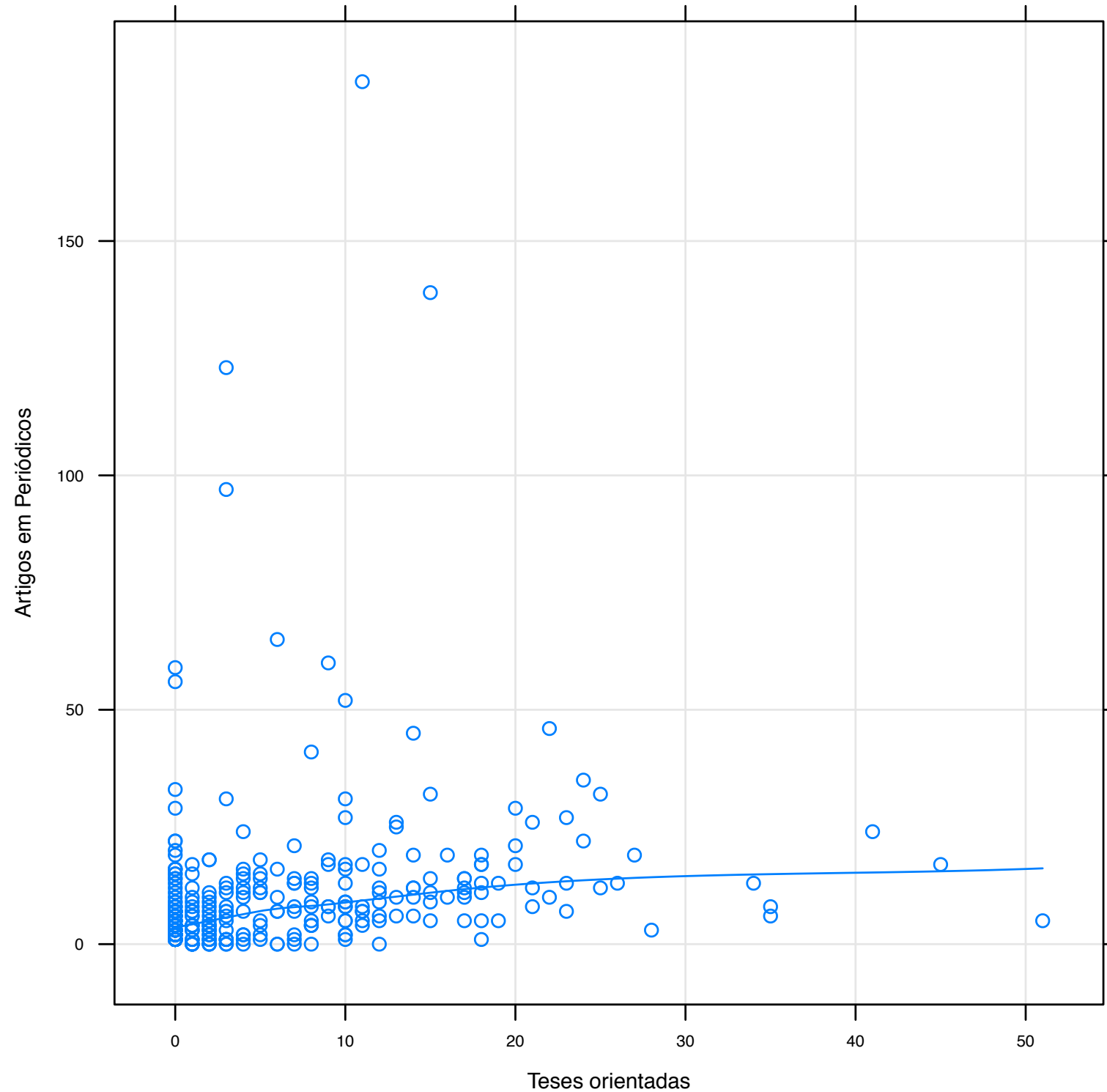
How old are we per department?



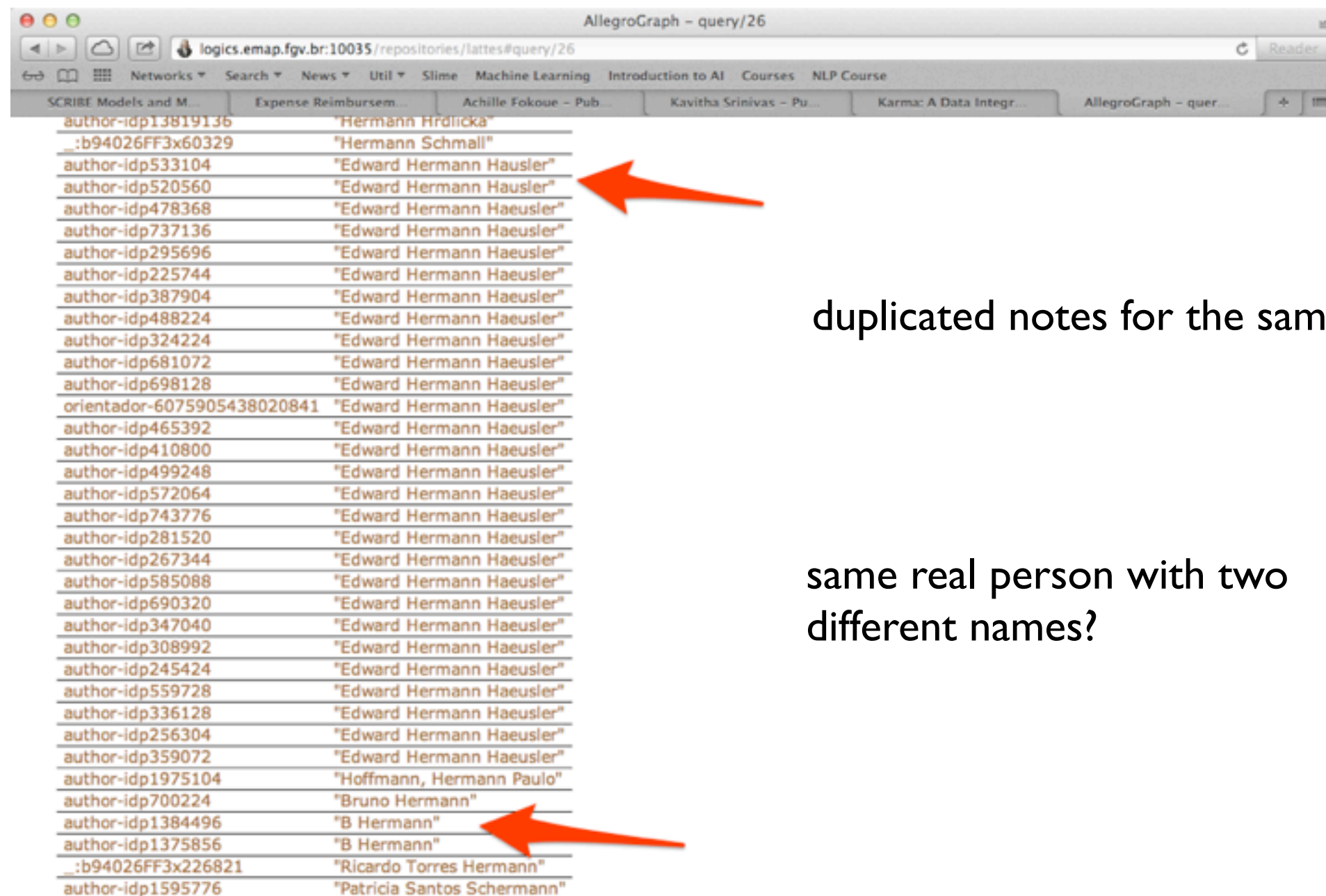
Publication quality?



Supervisions vs Publications



Data Problems



AllegroGraph - query/26

logics.emap.fgv.br:10035/repositories/lattes#query/26

Reader

Networks Search News Util Slime Machine Learning Introduction to AI Courses NLP Course

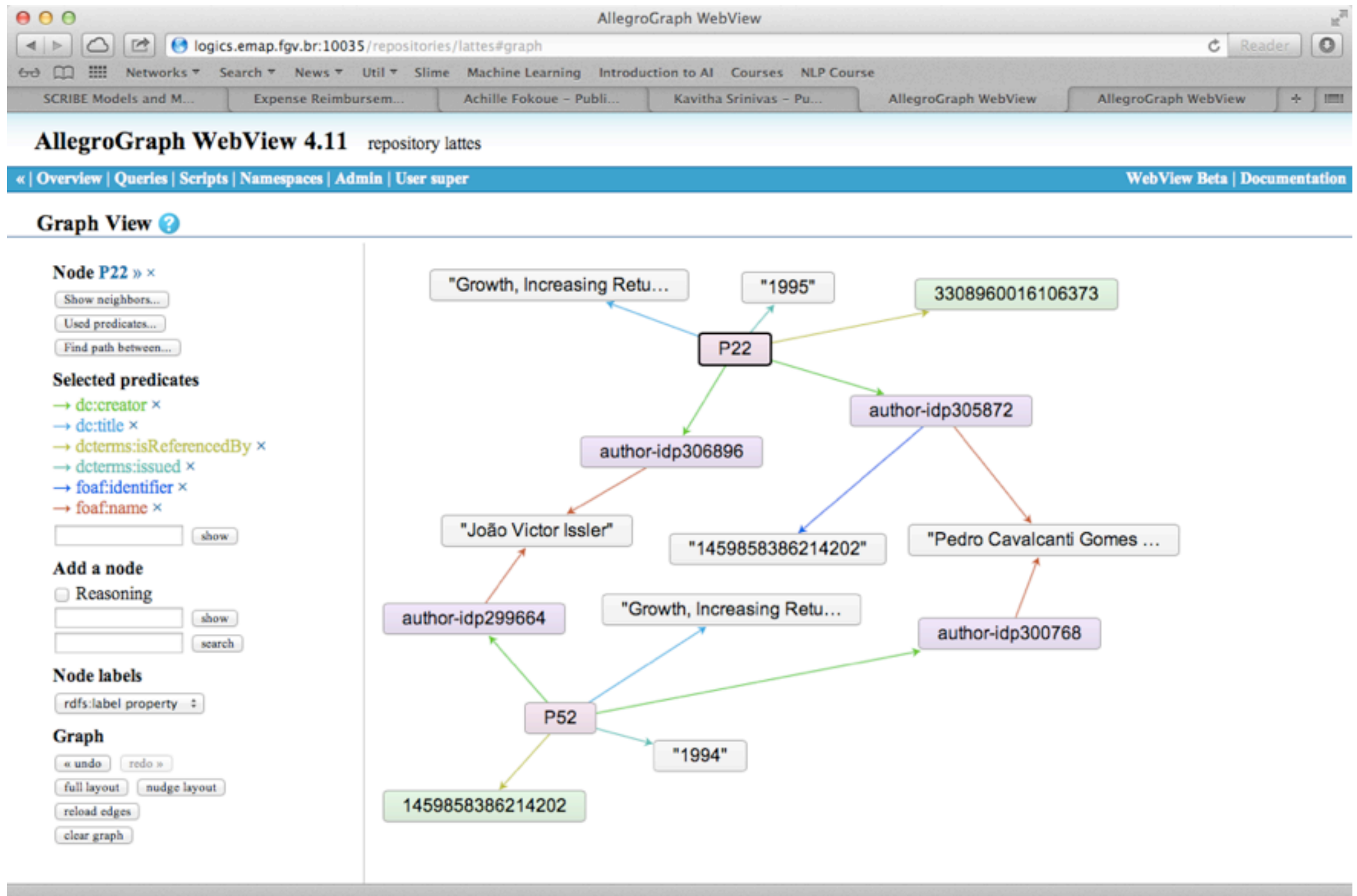
SCRIBE Models and M... Expense Reimbursem... Achille Fokoue - Pub... Kavitha Srinivas - Pu... Karma: A Data Integr... AllegroGraph - quer...

author-idp13819136	"Hermann Hrdlicka"
_:b94026FF3x60329	"Hermann Schmall"
author-idp533104	"Edward Hermann Hausler"
author-idp520560	"Edward Hermann Hausler"
author-idp478368	"Edward Hermann Haeusler"
author-idp737136	"Edward Hermann Haeusler"
author-idp295696	"Edward Hermann Haeusler"
author-idp225744	"Edward Hermann Haeusler"
author-idp387904	"Edward Hermann Haeusler"
author-idp488224	"Edward Hermann Haeusler"
author-idp324224	"Edward Hermann Haeusler"
author-idp681072	"Edward Hermann Haeusler"
author-idp698128	"Edward Hermann Haeusler"
orientador-6075905438020841	"Edward Hermann Haeusler"
author-idp465392	"Edward Hermann Haeusler"
author-idp410800	"Edward Hermann Haeusler"
author-idp499248	"Edward Hermann Haeusler"
author-idp572064	"Edward Hermann Haeusler"
author-idp743776	"Edward Hermann Haeusler"
author-idp281520	"Edward Hermann Haeusler"
author-idp267344	"Edward Hermann Haeusler"
author-idp585088	"Edward Hermann Haeusler"
author-idp690320	"Edward Hermann Haeusler"
author-idp347040	"Edward Hermann Haeusler"
author-idp308992	"Edward Hermann Haeusler"
author-idp245424	"Edward Hermann Haeusler"
author-idp559728	"Edward Hermann Haeusler"
author-idp336128	"Edward Hermann Haeusler"
author-idp256304	"Edward Hermann Haeusler"
author-idp359072	"Edward Hermann Haeusler"
author-idp1975104	"Hoffmann, Hermann Paulo"
author-idp700224	"Bruno Hermann"
author-idp1384496	"B Hermann"
author-idp1375856	"B Hermann"
_:b94026FF3x226821	"Ricardo Torres Hermann"
author-idp1595776	"Patricia Santos Schermann"

duplicated notes for the same entity

same real person with two different names?

Duplicated resources



Some duplication are easy to identify and remove!

The screenshot shows the AllegroGraph WebView 4.11 interface. The browser address bar displays `logics.emap.fgv.br:10035/repositories/lattes#query/29`. The page title is "AllegroGraph WebView 4.11 repository lattes". The navigation bar includes links for Overview, Queries, Scripts, Namespaces, Admin, and User super. The "Edit query" section shows the query language set to SPARQL, the query planner as default, and the result limit as 100. The query text is:

```
1 select ?t (count(?x) as ?tot) {  
2   ?x skos:prefLabel "CIENCIAS_SOCIAIS_APLICADAS" ;  
3   a ?t .  
4 }  
5 group by ?t
```

Below the query editor are buttons for "Execute", "Save", "Add to repository", and a link to "edit initfile". The "Result" section shows the query results downloaded as SPARQL JSON. The result is a table with two columns: "t" and "tot". The first row contains the value "fgvterms:grandeArea" for "t" and "19917" for "tot". A red arrow points to the value "19917" in the "tot" column.

t	tot
fgvterms:grandeArea	19917

Different sources and different descriptions

Advisor's Resume

Statements with **P725** » as the subject.

Predicate	Object	Context
dcterms:issued	"2010"	3386402716993689.rdf x
dc:title	"Políticas de ação afirmativa para negros no governo Fernando Henrique"	3386402716993689.rdf x
dc:creator	_:b94026FF3x104396	3386402716993689.rdf x
bibo:issuer	I000400000008	3386402716993689.rdf x
bibo:degree	bibo:degrees/ms	3386402716993689.rdf x
dcterms:isReferencedBy	3386402716993689	3386402716993689.rdf x
dc:contributor	mailto:mariacelina@daraujo.net	3386402716993689.rdf x
rdf:type	bibo:Thesis	3386402716993689.rdf x
dc:language	"Português"	3386402716993689.rdf x

Add statement...

Digital Library (DSpace)

Statements with **6891** as the subject.

Open a graph view on this node.

Predicate	Object	Context
dc:coverage	"http://bibliotecadigital.fgv.br/dspace/retrieve/17496/CPDOC2010MarilenedePaula.pdf.jpg"	records_hdl_10438_1758_1376495750.93.rdf x
dc:title	"Políticas de ação afirmativa para negros no governo Fernando Henrique Cardoso (1995-2002)"	records_hdl_10438_1758_1376495750.93.rdf x
dc:creator	"Paula, Marilene de"	records_hdl_10438_1758_1376495750.93.rdf x
dc:type	"Dissertation"	records_hdl_10438_1758_1376495750.93.rdf x
dc:subject	"Política governamental"	records_hdl_10438_1758_1376495750.93.rdf x
dc:subject	"Discriminação racial"	records_hdl_10438_1758_1376495750.93.rdf x
dc:subject	"Programas de ação afirmativa"	records_hdl_10438_1758_1376495750.93.rdf x
dc:subject	"Relações raciais"	records_hdl_10438_1758_1376495750.93.rdf x
dc:identifier	"http://hdl.handle.net/10438/6891"	records_hdl_10438_1758_1376495750.93.rdf x
dc:date	"2010-07-19T12:34:26Z"	records_hdl_10438_1758_1376495750.93.rdf x
dc:date	"2010-04-08"	records_hdl_10438_1758_1376495750.93.rdf x
dc:contributor	"Pandolfi, Dulce Chaves"	records_hdl_10438_1758_1376495750.93.rdf x
dc:contributor	"Feres Junior, João"	records_hdl_10438_1758_1376495750.93.rdf x
dc:contributor	"D'Araujo, Maria Celina Soares"	records_hdl_10438_1758_1376495750.93.rdf x
dc:contributor	"Freire-Medeiros, Bianca"	records_hdl_10438_1758_1376495750.93.rdf x
rdf:type	ontobib:Publication	records_hdl_10438_1758_1376495750.93.rdf x
dc:language	"pt_BR"	records_hdl_10438_1758_1376495750.93.rdf x
dc:description	"O presente trabalho busca, através da análise da conjuntura política das relações raciais no Brasil no final dos anos 1990 e início dos anos 2000, apontar para o surgimento de um campo específico das políticas públicas: a promoção da igualdade racial. Para obter tal finalidade analiso a trajetória das políticas de ação afirmativa do governo Fernando Henrique Cardoso (1995-2002), desenvolvidas em vários Ministérios, tais como Justiça, Desenvolvimento Agrário, Educação, Relações Exteriores e Trabalho e também no Supremo Tribunal Federal para determinar qual a contribuição e significado de tais políticas e do discurso político-simbólico desse governo para o avanço do debate sobre as relações raciais no Brasil."	records_hdl_10438_1758_1376495750.93.rdf x
dc:source	"CPDOC - Centro de Pesquisa e Documentação de História Contemporânea do Brasil"	records_hdl_10438_1758_1376495750.93.rdf x
dc:source	"CPDOC - Dissertações, Mestrado Profissional em Bens Culturais e Projetos Sociais"	records_hdl_10438_1758_1376495750.93.rdf x

Add statement...

source

But what <http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements#terms-contributor> says?

No reliable IDs from Lattes!

Statements with **author-idp2221456** » as the subject.

Predicate	Object	Context	
foaf:citationName	"LEONI, E."	1959285307720446.rdf	×
foaf:name	"Eduardo Leoni"	1959285307720446.rdf	×
rdf:type	foaf:Agent	1959285307720446.rdf	×
foaf:identifier	"1558145083566407"	1959285307720446.rdf	×

[Add statement...](#)

Statements with **author-idp2221456** » as the predicate.

[Add statement...](#)

Statements with **author-idp2221456** » as the object.

Subject	Predicate	Context	
P53	dc:creator	1959285307720446.rdf	×
_:bC3F14E04x80871	rdf:first	1959285307720446.rdf	×

[Add statement...](#)

Statements with **author-idp1842688** » as the subject.

Predicate	Object	Context	
foaf:citationName	"ALSTON, Lee"	1959285307720446.rdf	×
foaf:name	"Lee Alston"	1959285307720446.rdf	×
rdf:type	foaf:Agent	1959285307720446.rdf	×
foaf:identifier	"1558145083566407"	1959285307720446.rdf	×

[Add statement...](#)

Statements with **author-idp1842688** » as the predicate.

[Add statement...](#)

Statements with **author-idp1842688** » as the object.

Subject	Predicate	Context	
P163	dc:creator	1959285307720446.rdf	×
_:bC3F14E04x80558	rdf:first	1959285307720446.rdf	×

[Add statement...](#)

```
NOME-DA-EDITORIA="" CIDADE-DA-EDITORIA="" />
<AUTORES NOME-COMPLETO-DO-AUTOR="Carlos Eduardo Ferreira Pereira Filho"
NOME-PARA-CITACAO="PEREIRA, C." ORDEM-DE-AUTORIA="1"
NRO-ID-CNPQ="" />
<AUTORES NOME-COMPLETO-DO-AUTOR="Bernardo Mueller"
NOME-PARA-CITACAO="MUELLER, Bernardo" ORDEM-DE-AUTORIA="2"
NRO-ID-CNPQ="1558145083566407" />
<PALAVRAS-CHAVE PALAVRA-CHAVE-1="Committee power"
PALAVRA-CHAVE-2="positive theory"
PALAVRA-CHAVE-3="rational choice" PALAVRA-CHAVE-4=""
PALAVRA-CHAVE-5="" PALAVRA-CHAVE-6="" />
<AREAS-DO-CONHECIMENTO>
<AREA-DO-CONHECIMENTO-1 NOME-GRANDE-AREA-DO-CONHECIMENTO="CIENCIAS_HUM
NOME-DA-AREA-DO-CONHECIMENTO="Ci&#234;ência Pol&#237;ítica"
NOME-DA-SUB-AREA-DO-CONHECIMENTO="Comportamento Pol&#237;ítico"
NOME-DA-ESPECIALIDADE="Comportamento Legislativo" />
<AREA-DO-CONHECIMENTO-2 NOME-GRANDE-AREA-DO-CONHECIMENTO="CIENCIAS HUM
```


Bad data input!

```
<TRABALHO-EM-EVENTOS SEQUENCIA-PRODUCAO="259">
  <DADOS-BASICOS-DO-TRABALHO NATUREZA="COMPLETO"
  TITULO-DO-TRABALHO="Gest&#227;o do Sistema Produtivo e Inova&#231;&#227;o Tecnol&#243;gica"
  ANO-DO-TRABALHO="1993" PAIS-DO-EVENTO="Brasil"
  IDIOMA="Portugu&#234;s" MEIO-DE-DIVULGACAO="IMPRESS&#227;o"
  HOME-PAGE-DO-TRABALHO="" FLAG-RELEVANCIA="NAO" DOI=""
  TITULO-DO-TRABALHO-INGLES=""
  FLAG-DIVULGACAO-CIENTIFICA="NAO" />
  <DETALHAMENTO-DO-TRABALHO CLASSIFICACAO-DO-EVENTO="INTERNACIONAL"
  NOME-DO-EVENTO="V Semin&#225;rio Latinoamericano de Gest&#243;n Tecnol&#243;gica"
  CIDADE-DO-EVENTO="Bogot&#225;/Colombia"
  ANO-DE-REALIZACAO="1993"
  TITULO-DOS-ANAIS-OU-PROCEEDINGS="V Semin&#225;rio Latinoamericano de Gest&#243;n Tecnol&#243;gica"
  VOLUME="1" FASCICULO="1" SERIE="" PAGINA-INICIAL="13"
  PAGINA-FINAL="21" ISBN=""
  NOME-DA-EDITORIA="Asociaci&#243;n Latinoamericana de Gest&#243;n Tecnol&#243;gica"
  CIDADE-DA-EDITORIA="Bogot&#225;" />
  <AUTORES NOME-COMPLETO-DO-AUTOR="Jose Carlos Barbieri"
  NOME-PARA-CITACAO="BARBIERI, Jose Carlos;BARBIERI, JOS&#201; CARLOS"
  ORDEM-DE-AUTORIA="1" />
  <AREAS-DO-CONHECIMENTO>
    <AREA-DO-CONHECIMENTO-1 NOME-GRANDE-AREA-DO-CONHECIMENTO="CIENCIAS_SOCIAIS_APLICADAS"
    NOME-DA-AREA-DO-CONHECIMENTO="Administra&#231;&#227;o"
    NOME-DA-SUB-AREA-DO-CONHECIMENTO="Administra&#231;&#227;o de Setores Espec&#237;ficos"
    NOME-DA-ESPECIALIDADE="" />
  </AREAS-DO-CONHECIMENTO>
  <INFORMACOES-ADICIONAIS DESCRICAO-INFORMACOES-ADICIONAIS=""
  DESCRICAO-INFORMACOES-ADICIONAIS-INGLES="" />
</TRABALHO-EM-EVENTOS>
```

Statements with **_:b94026FF3x304** as the subject. Open a graph

Predicate	Object	Context
gn:name	"Bogotá/Colombia"	0041377800166678.rdf ×
rdf:type	SpatialThing	0041377800166678.rdf ×
gn:countrycode	"Brasil"	0041377800166678.rdf ×

[Add statement...](#)

Statements with **_:b94026FF3x304** as the predicate.

[Add statement...](#)

Statements with **_:b94026FF3x304** as the object.

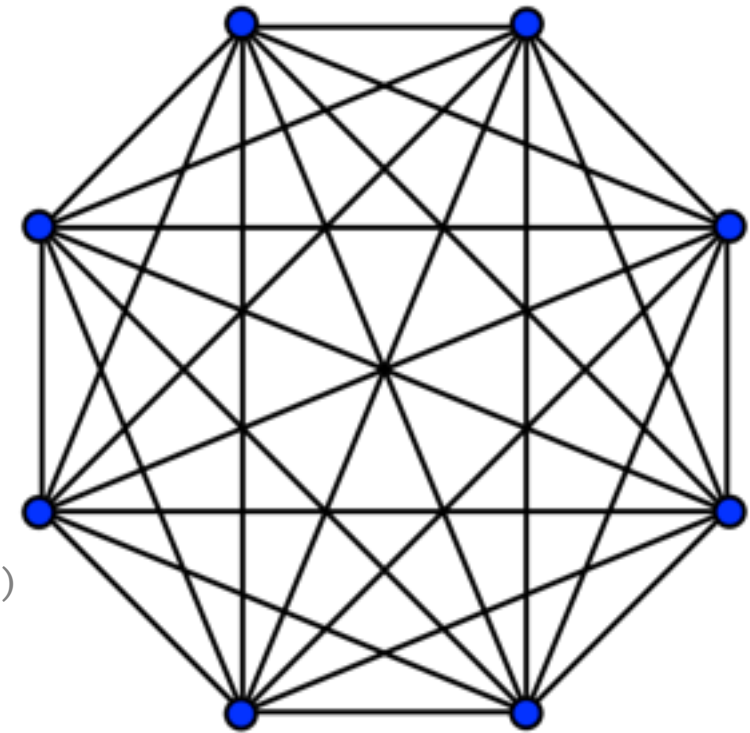
Subject	Predicate	Context
_:b94026FF3x303	event:place	0041377800166678.rdf ×

[Add statement...](#)

ad-hoc deduplication

```
(defun assert-same-list (list)
  (let ((new nil))
    (mapcar (lambda (pair)
              (let ((a (first pair))
                    (b (second pair)))
                (if (not (blank-node-p a))
                    (push (reverse pair) new)
                    (push pair new)))))
      list)
  (dolist (pair new)
    (add-triple (first pair) !owl:sameAs (second pair)))))
```

```
(select0/callback (?x ?y) #'insert-same-as
  (q- ?x !rdf:type !foaf:Agent)
  (q- ?y !rdf:type !foaf:Agent)
  (q- ?x !foaf:name ?n)
  (q- ?y !foaf:name ?n)
  (lispp (upi< ?x ?y)))
```



Naive approach: Shaking hands!

ad-hoc deduplication

```
(defun components (vertices n generator)
  (do ((res nil)
      (vtx vertices
          (set-difference vtx (car res) :test #'upi=)))
    ((null vtx) res)
    (push (ego-group (car vtx) n generator) res)))

(defsna-generator same-journal (node)
  (select0 (?j)
    (q- (?? node) !bibo:issn ?i)
    (q- ?j !bibo:issn ?i)
    (lispp (utils::check-issn (part->value ?i)))
    (lispp (upi< node ?j))
    (q- ?j !dc:title ?t2)
    (q- (?? node) !dc:title ?t1)
    (lispp (> (utils::jaro-winkler-distance (part->value ?t1) (part->value ?t2)) 0.7))))

(let ((nodes (mapcar #'subject (get-triples-list :p !bibo:issn :limit nil))))
  (dolist (g (components nodes 2 'same-journal))
    (merge-nodes g)))
```

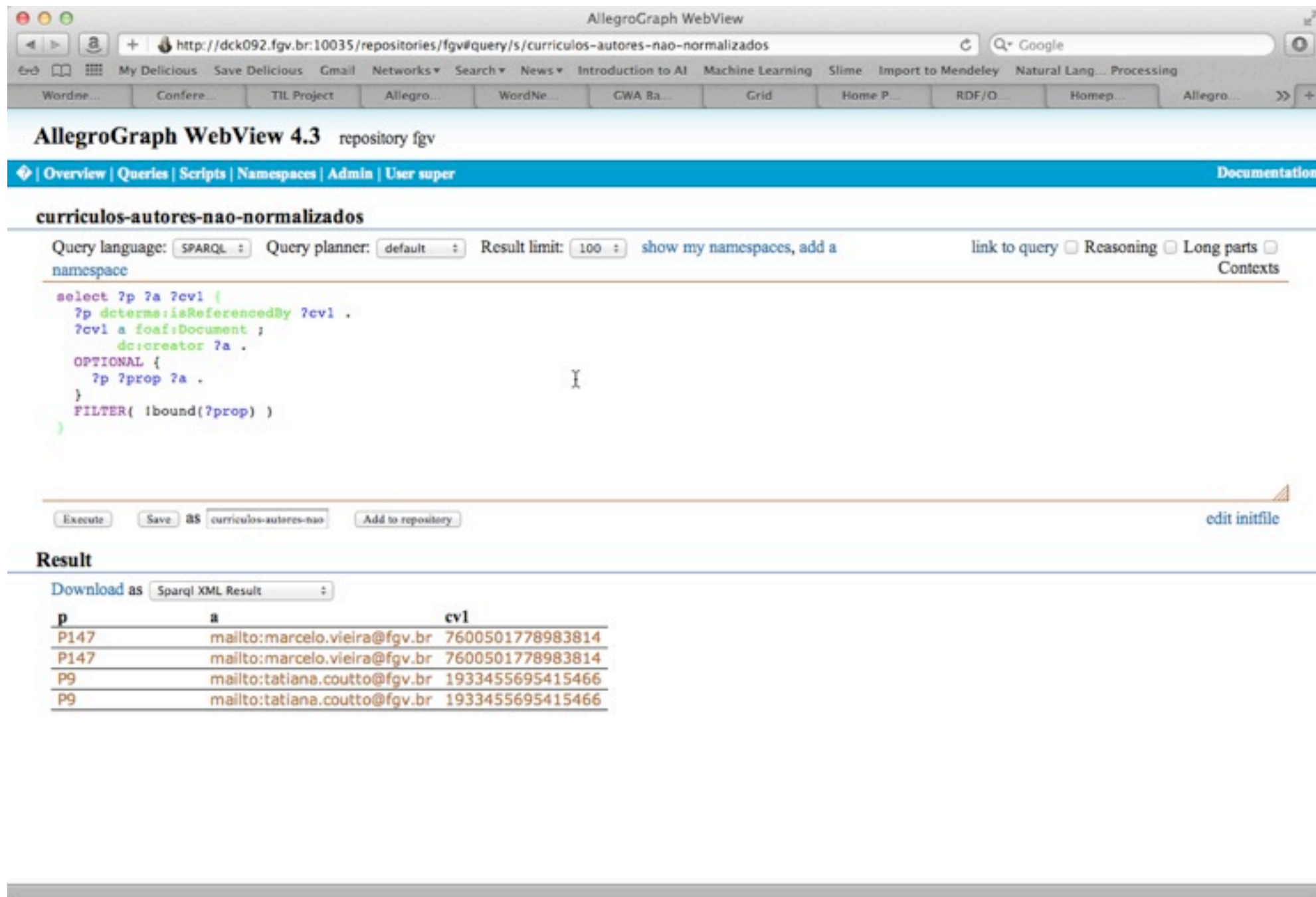
An ad-hoc solution: breath-first-search of connected components!

How to deal with those data quality problems?

- ~750 CV Lattes and collected data from other sources (Digital Library etc) in one triple store.
- lots of errors (inconsistencies) for different reasons: poor user interface for input data, misinterpretation etc.
- How to identify the errors? (non *ad-hoc* matter)
- How to fix what can be fixed automatically? Sources reputations and propagation of reputations!
- Ongoing research!
- Pellet Integrity Constraints: Validating RDF with OWL. (<http://clarkparsia.com/pellet/icv/>)
- **Truth Maintenance!** Integrity enforcement! Partial repairs! DB researches.

Query as constraints:

An article referenced by a CV must have the author of this CV as one of its authors!



The screenshot shows the AllegroGraph WebView 4.3 interface. The browser address bar displays the URL: `http://dck092.fgv.br:10035/repositories/fgv#query/s/curriculos-autores-nao-normalizados`. The page title is "AllegroGraph WebView 4.3 repository fgv". The navigation bar includes links for Overview, Queries, Scripts, Namespaces, Admin, and User super. The main content area shows the query language set to SPARQL, the query planner set to default, and the result limit set to 100. The query is as follows:

```
select ?p ?a ?cvl {
  ?p dcterms:isReferencedBy ?cvl .
  ?cvl a foaf:Document ;
       dc:creator ?a .
  OPTIONAL {
    ?p ?prop ?a .
  }
  FILTER( !bound(?prop) )
}
```

Below the query, there are buttons for Execute, Save, and Add to repository. The result section shows the results in a table format, with columns p, a, and cvl. The results are as follows:

p	a	cvl
P147	mailto:marcelo.vieira@fgv.br	7600501778983814
P147	mailto:marcelo.vieira@fgv.br	7600501778983814
P9	mailto:tatiana.coutto@fgv.br	1933455695415466
P9	mailto:tatiana.coutto@fgv.br	1933455695415466

Query as constraints:

If two resources were identified as being the same article (same title), every author of the first one should also be author of the second one!

The screenshot shows the AllegroGraph WebView interface. The browser address bar displays the URL `http://dck092.fgv.br:10035/repositories/fgv#query/3`. The page title is "AllegroGraph WebView 4.3 repository fgv". The navigation bar includes links for Overview, Queries, Scripts, Namespaces, Admin, and User super. The "Edit query" section shows the query language set to SPARQL, the query planner set to default, and the result limit set to 100. The query text is as follows:

```
select ?p1 ?p2 {
  ?p1 dc:title ?t ;
      dc:creator ?c .
  ?p2 dc:title ?t .
  OPTIONAL {
    ?p2 ?rel ?c .
  }
  FILTER( !bound(?rel) )
}
```

Below the query editor, there are buttons for "Execute", "Save", and "Add to repository". The "Result" section shows the download format set to "Sparql XML Result". The results are displayed in a table with two columns, p1 and p2.

p1	p2
P1000	_:b40948D10x122821
P79	P93
P19	P32
P8	_:b40948D10x7408
P8	_:b40948D10x39947
P88	P588
P86	P538
P27	P32
P26	P285
P72	P504
P93	P540
P125	P647
P1000	_:b40948D10x122821

But of course title is not enough. Refining last example:

Of course, two publications cannot be considered the same comparing only their titles!

We need entity alignment, similarity checker...

Suppose we have identified all resources that represent the same real “entity” using owl:sameAs, than ...

```
ask {  
  ?p1 owl:sameAs ?p2 ;  
      dc:creator ?c .  
  OPTIONAL {  
    ?p2 ?rel ?c .  
  }  
  FILTER( !bound(?rel) )  
}
```

Next Steps

- Focus (research opportunities):
 - data normalization and cleanup (results from DB researches)
 - ontologies alignment and instances matching
- Web interface for browsing and queries (dev, but important):
 - RDF to Solr and HTML/JS with Solr backend
 - **Use VIVO (opportunities: network of installations, ontology alignment)**
 - push to <https://www.researchgate.net/>
 - Use <http://bibapp.org>