



UNIVERSITÀ  
DEGLI STUDI DI TRIESTE

**DIPARTIMENTO DI INGEGNERIA E ARCHITETTURA**

**Corso di Laurea Triennale in Ingegneria dell'Informazione**

# **ANALISI DELLE FUNZIONALITÀ DI INDICIZZAZIONE E RICERCA DOCUMENTI DELLA PIATTAFORMA APACHE SOLR**

**Relatore**

*Prof. Alberto Bartoli*

**Correlatore**

*Prof. Eric Medvet*

**Laureando**

*Livio Bisogni*



# Descrizione del problema

- a) Configurare in forma prototipale Apache Solr (motore di ricerca per documenti locali)
- b) Data una cartella:
  - 1. Individuare in tempo reale le modifiche ai file in essa presenti
  - 2. Identificare i file pdf non ancora indicizzabili per contenuto
  - 3. Rendere tali file indicizzabili per contenuto
  - 4. Indicizzare i file su un motore di ricerca (Solr)
  - 5. Effettuare delle ricerche (possibilmente con interfaccia grafica)



# Descrizione del problema

- a) Configurare in forma prototipale Apache Solr (motore di ricerca per documenti locali)
- b) Data una cartella:
  - 1. Individuare in tempo reale le modifiche ai file in essa presenti  
← *Script*
  - 2. Identificare i file pdf non ancora indicizzabili per contenuto
  - 3. Rendere tali file indicizzabili per contenuto
  - 4. Indicizzare i file su un motore di ricerca (Solr)
  - 5. Effettuare delle ricerche (possibilmente con interfaccia grafica)

# Descrizione del problema

a) Configurare in forma prototipale Apache Solr (motore di ricerca per documenti locali)

b) Data una cartella:

1. Individuare in tempo reale le modifiche ai file in essa presenti  
← *Script*

2. Identificare i file pdf non ancora indicizzabili per contenuto  
← *Programma Java*

3. Rendere tali file indicizzabili per contenuto

4. Indicizzare i file su un motore di ricerca (Solr)

5. Effettuare delle ricerche (possibilmente con interfaccia grafica)

# Descrizione del problema

a) Configurare in forma prototipale Apache Solr (motore di ricerca per documenti locali)

b) Data una cartella:

1. Individuare in tempo reale le modifiche ai file in essa presenti  
← *Script*

2. Identificare i file pdf non ancora indicizzabili per contenuto  
← *Programma Java*

3. Rendere tali file indicizzabili per contenuto  
← *Script*

4. Indicizzare i file su un motore di ricerca (Solr)

5. Effettuare delle ricerche (possibilmente con interfaccia grafica)

# Descrizione del problema

a) Configurare in forma prototipale Apache Solr (motore di ricerca per documenti locali)

b) Data una cartella:

1. Individuare in tempo reale le modifiche ai file in essa presenti  
← *Script*

2. Identificare i file pdf non ancora indicizzabili per contenuto  
← *Programma Java*

3. Rendere tali file indicizzabili per contenuto  
← *Script*

4. Indicizzare i file su un motore di ricerca (Solr)  
← *Script*

5. Effettuare delle ricerche (possibilmente con interfaccia grafica)

# Descrizione del problema

a) Configurare in forma prototipale Apache Solr (motore di ricerca per documenti locali)

b) Data una cartella:

1. Individuare in tempo reale le modifiche ai file in essa presenti  
← *Script*

2. Identificare i file pdf non ancora indicizzabili per contenuto  
← *Programma Java*

3. Rendere tali file indicizzabili per contenuto  
← *Script*

4. Indicizzare i file su un motore di ricerca (Solr)  
← *Script*

5. Effettuare delle ricerche (possibilmente con interfaccia grafica)  
← *UI Solritas*

# Apache Solr

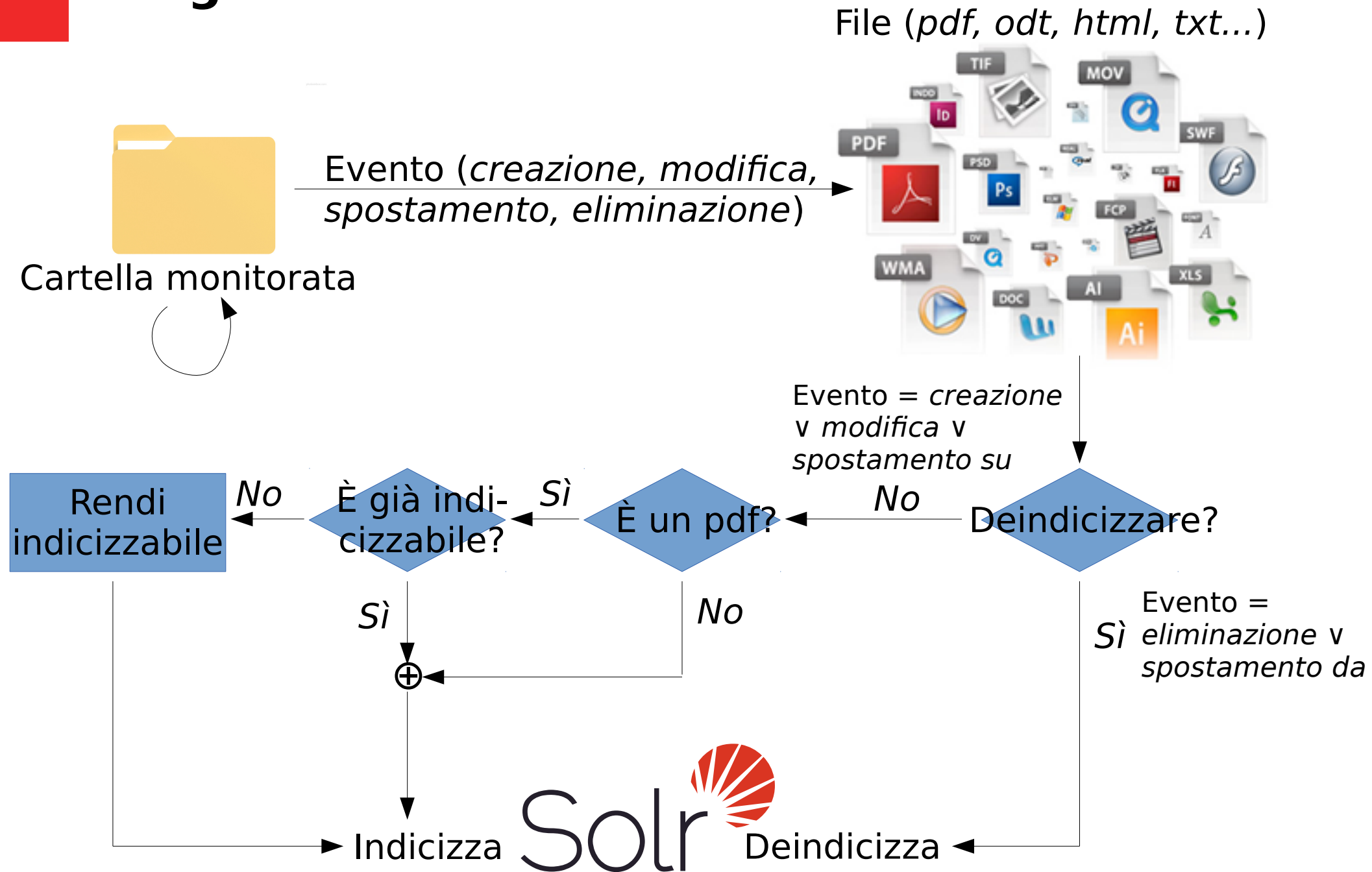


- È una piattaforma di indicizzazione e ricerca di documenti open source nata nel 2004





# Diagramma di flusso



# Configurazione

**HOST**

*Cartella*



*condivisa*

**GUEST**

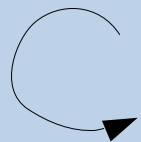


# Configurazione

## HOST

**1**

Script per  
monito-  
raggio  
modifiche  
cartella  
condivisa  
(*inotify*)

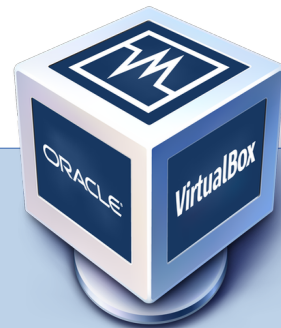


*Cartella*



*condivisa*

## GUEST

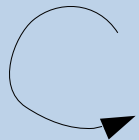


# Configurazione

## HOST

**1**

Script per  
monito-  
raggio  
modifiche  
cartella  
condivisa  
(*inotify*)



*Cartella*



*condivisa*

## GUEST



**2**

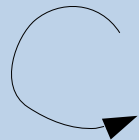
Programma Java per  
identificazione pdf non  
ancora indicizzabili

# Configurazione

## HOST

**1**

Script per  
monito-  
raggio  
modifiche  
cartella  
condivisa  
(*inotify*)



*Cartella*



*condivisa*

## GUEST



**2**

Programma Java per  
identificazione pdf non  
ancora indicizzabili

**3** Script per conversione  
OCR (*tesseract*)

# Configurazione

```
SimplePostTool version 5.0.0
Posting files to [base] url http://localhost:8983/solr/files/update...
Entering auto mode. File endings considered are xml,json,jsonl,csv,pdf,doc,docx,
ppt,pptx,xls,xlsx,odt,odp,ods,ott,otp,ots,rtf,htm,html,txt,log
POSTing file 001034919-untscle-a60f60ce-392b-43dc-alac-214b25ee6537-000.pdf (app
lication/pdf) to [base]/extract
1 files indexed.
COMMITting Solr index changes to http://localhost:8983/solr/files/update...
Time spent: 0:00:00.578
```

HTTP POST REQUEST

**HOST**

**1**

Script per  
monito-  
raggio  
modifiche  
cartella  
condivisa  
(*inotify*)

*Cartella*



*condivisa*

**4**  
INDICIZZAZIONE/  
DEINDICIZZAZIONE

**GUEST**

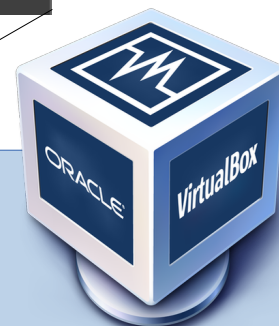
Solr



**2**

Programma Java per  
identificazione pdf non  
ancora indicizzabili

**3** Script per conversione  
OCR (*tesseract*)



# Configurazione

```
SimplePostTool version 5.0.0
Posting files to [base] url http://localhost:8983/solr/files/update...
Entering auto mode. File endings considered are xml,json,jsonl,csv,pdf,doc,docx,
ppt,pptx,xls,xlsx,odt,odp,ods,ott,otp,ots,rtf,htm,html,txt,log
POSTing file 001034919-untscl-a60f60ce-392b-43dc-alac-214b25ee6537-000.pdf (app
lication/pdf) to [base]/extract
1 files indexed.
COMMITting Solr index changes to http://localhost:8983/solr/files/update...
Time spent: 0:00:00.578
```

HTTP POST REQUEST

HOST

1

Script per  
monito-  
raggio  
modifiche  
cartella  
condivisa  
(*inotify*)

Cartella



condivisa

INDICIZZAZIONE/  
DEINDICIZZAZIONE

QUERY

5

HTTP GET REQUEST

GUEST

Solr

2

Programma Java per  
identificazione pdf non  
ancora indicizzabili

3 Script per conversione  
OCR (*tesseract*)



```
v@v-PC ~ $ wget "http://localhost:8983/solr/files/browse?fq=language%3A%22it%22&q=dipartimento+universitario+clinico&type=pdf"
--2016-10-03 16:37:10-- http://localhost:8983/solr/files/browse?fq=language%3A%22it%22&q=dipartimento+universitario+clinico&type=pdf
Resolving localhost (localhost)... 127.0.0.1
Connecting to localhost (localhost)|127.0.0.1|:8983... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: 'browse?fq=language%3A%22it%22&q=dipartimento+universitario+clinico&type=pdf'

browse?fq=language%3A%22it%22&q=di      [ <=> ] 227,57K  ---KB/s   in 0,05s

2016-10-03 16:37:10 (4,67 MB/s) - 'browse?fq=language%3A%22it%22&q=dipartimento+universitario+clinico&type=pdf' saved [233033]
```

# Apache Tika



- È un framework per il riconoscimento e l'estrazione di testo e metadati („dati che forniscono informazioni su altri dati“)
- Supporta file di molte estensioni:
  - *HyperText Markup Language (HTML, ...)*
  - *XML and derived formats (XML, XHTML, ...)*
  - *Microsoft Office document formats (DOC, DOCX, PPT, PPTX, XLS, ...)*
  - *OpenDocument Format (ODF, ODT, ODS, ODP, ODG, ODF, ...)*
  - *Portable Document Format (PDF, ...)*
  - *Electronic Publication Format (EPUB, IBOOKS, ...)*
  - *Compression and packaging formats (ZIP, GZIP, TAR, 7Z, RAR...)*
  - *Text formats (TXT, ...)*
  - *Image formats (JPEG, TIFF, BMP, PNG, GIF, ...)*
  - *Audio formats (MP3, WAV, AIFF, FLAC MIDI, ...)*
  - *Video formats (MP4, AVI, MPEG, FLV, ...)*
  - *...e moltissimi altri formati*





# UI Solritas (I)


- È l'interfaccia utente di ricerca già integrata in Solr


UNIVERSITÀ  
DEGLI STUDI DI TRIESTE


Cerca:


121 risultati trovati in 21ms Pagina 1 di 13

[Tutti i formati \(121\)](#) [PDF \(114\)](#) [File testuali \(3\)](#) [HTML \(1\)](#) [Presentazioni \(1\)](#) [Sconosciuto \(2\)](#)

 **Cose da fare per HDP.odt**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/Cose da fare per HDP.odt

 **000971031-untscle-64a0e436-8295-440f-8b31-522119994711-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/000971031-untscle-64a0e436-8295-440f-8b31-522119994711-000.pdf

 **Titolo Classe**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/000936617-untscle-8668f761-aac9-46ce-bbe1-2edc4b734579-000.pdf

 **000963647-untscle-1d52cd86-3078-4012-b287-04dc3f5d568e-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/000963647-untscle-1d52cd86-3078-4012-b287-04dc3f5d568e-000.pdf

# UI Solritas (II)

## HTTP GET REQUEST

```
v@v-PC ~ $ wget "http://localhost:8983/solr/files/browse?fq=language%3A%22it%22&q=dipartimento+universitario+clinico&type=pdf"
--2016-10-03 16:37:10-- http://localhost:8983/solr/files/browse?fq=language%3A%22it%22&q=dipartimento+universitario+clinico&type=pdf
Resolving localhost (localhost)... 127.0.0.1
Connecting to localhost (localhost)|127.0.0.1|:8983... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: 'browse?fq=language%3A%22it%22&q=dipartimento+universitario+clinico&type=pdf'

browse?fq=language%3A%22it%22&q=di      [ <=> ] 227,57K  ---KB/s   in 0,05s

2016-10-03 16:37:10 (4,67 MB/s) - 'browse?fq=language%3A%22it%22&q=dipartimento+universitario+clinico&type=pdf' saved [233033]
```

Solr browse: files - Mozilla Firefox

Solr browse: files x +

localhost:8983/solr/files/browse?q=dipartimento+universitario+clinico&type=pdf&fq=language%3A"it" Cerca


UNIVERSITÀ  
DEGLI STUDI DI TRIESTE

Cerca:  Cerca

> language:"it" [x](#)

30 risultati trovati in 44ms Pagina 1 di 3

Tutti i formati (30) PDF (30) File testuali (0)

 001026124-untscl-7bee8dc3-c8d9-49dc-b7ca-1c3e682cebf3-000.pdf

id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001026124-untscl-7bee8dc3-c8d9-49dc-b7ca-1c3e682cebf3-000.pdf

.2015t2016 *Dipartimento Universitario Clinico* di Scienze Mediche, Ghirurgiche e della Salute Si rende

# UI Solritas (III)

- Oltre alla mera ricerca, molte possibilità:

Solr browse: files - Mozilla Firefox

Solr browse: files

localhost:8983/solr/files/browse?q=dipartimento+universitario+clinico&type=pdf&fq=language%3A%22it%22

Cerca

UNIVERSITÀ DEGLI STUDI DI TRIESTE

Cerca:  Cerca

> language:"it" [x](#)

30 risultati trovati in 44ms Pagina 1 di 3

Tutti i formati (30) PDF (30) File testuali (0)

**001026124-untscl-7bee8dc3-c8d9-49dc-b7ca-1c3e682cebf3-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001026124-untscl-7bee8dc3-c8d9-49dc-b7ca-1c3e682cebf3-000.pdf  
.201512016 *Dipartimento Universitario Clinico* di Scienze Mediche, Ghirurgiche e della Salute Si rende

**001018758-untscl-a4051605-0159-4471-9fc6-1a03b52fadfb-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001018758-untscl-a4051605-0159-4471-9fc6-1a03b52fadfb-000.pdf  
PREMIO DI STUDIO IN MEMORIA DI LAURA VITTORI Anno Accademico 201 512016 *Dipartimento Universitario*

**BOZZA nuova Copertina modello informazione rischi 1.2 CORRETTA SESSIMI**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001032790-untscl-8dc5cc64-9ce3-4614-8fbf-6d339055de9a-003.pdf  
Salute Via Monte Cengio (Palestra C.U.S.) Via Manzoni 16 - *Dipartimento Universitario Clinico* di

**001032229-untscl-e64969b4-ca2e-4812-85ac-97a3fe12c709-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001032229-untscl-e64969b4-ca2e-4812-85ac-97a3fe12c709-000.pdf  
del Consiglio del *Dipartimento Universitario Clinico* di Scienze mediche, chirurgiche e della salute

**01 DOMANDA DI ACCREDITAMENTO FOTOGRAFI 2016 - IMPRESE**

**Frase chiave**  
corso di validità degli studi di  
del del d.p.r di un documento  
il dipartimento di in caso di  
in corso di legge studi di trieste

**Lingua**  
[Italian](#) (30)

# UI Solritas (III)

- Oltre alla mera ricerca, molte possibilità:

Solr browse: files - Mozilla Firefox

Solr browse: files

localhost:8983/solr/files/browse?q=dipartimento+universitario+clinico&type=pdf&fq=language%3A"it"

Cerca

Lingua di navigazione utente →

Cerca:  Cerca

> language:"it" [x](#)

30 risultati trovati in 44ms Pagina 1 di 3

Tutti i formati (30) PDF (30) File testuali (0)

**001026124-untscl-7bee8dc3-c8d9-49dc-b7ca-1c3e682cebf3-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001026124-untscl-7bee8dc3-c8d9-49dc-b7ca-1c3e682cebf3-000.pdf  
.201512016 *Dipartimento Universitario Clinico* di Scienze Mediche, Ghirurgiche e della Salute Si rende

**001018758-untscl-a4051605-0159-4471-9fc6-1a03b52fadfb-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001018758-untscl-a4051605-0159-4471-9fc6-1a03b52fadfb-000.pdf  
PREMIO DI STUDIO IN MEMORIA DI LAURA VITTORI Anno Accademico 201 512016 *Dipartimento Universitario*

**BOZZA nuova Copertina modello informazione rischi 1.2 CORRETTA SESSIMI**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001032790-untscl-8dc5cc64-9ce3-4614-8fbf-6d339055de9a-003.pdf  
Salute Via Monte Cengio (Palestra C.U.S.) Via Manzoni 16 - *Dipartimento Universitario Clinico* di

**001032229-untscl-e64969b4-ca2e-4812-85ac-97a3fe12c709-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001032229-untscl-e64969b4-ca2e-4812-85ac-97a3fe12c709-000.pdf  
del Consiglio del *Dipartimento Universitario Clinico* di Scienze mediche, chirurgiche e della salute

**01 DOMANDA DI ACCREDITAMENTO FOTOGRAFI 2016 - IMPRESE**

**Frase chiave**  
corso di validità degli studi di  
del del d.p.r di un documento  
il dipartimento di in caso di  
in corso di legge studi di trieste

**Lingua**  
[Italian](#) (30)

# UI Solritas (III)

- Oltre alla mera ricerca, molte possibilità:

Solr browse: files - Mozilla Firefox

localhost:8983/solr/files/browse?q=dipartimento+universitario+clinico&type=pdf&fq=language%3A"it"

Cerca

Lingua di navigazione utente →

Cerca:  Cerca

> language:"it" [x](#)

30 risultati trovati in 44ms Pagina 1 di 3

Tutti i formati (30) PDF (30) File testuali (0)

**001026124-untscl-7bee8dc3-c8d9-49dc-b7ca-1c3e682cebf3-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001026124-untscl-7bee8dc3-c8d9-49dc-b7ca-1c3e682cebf3-000.pdf  
.201512016 *Dipartimento Universitario Clinico* di Scienze Mediche, Ghirurgiche e della Salute Si rende

**001018758-untscl-a4051605-0159-4471-9fc6-1a03b52fadfb-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001018758-untscl-a4051605-0159-4471-9fc6-1a03b52fadfb-000.pdf  
PREMIO DI STUDIO IN MEMORIA DI LAURA VITTORI Anno Accademico 201 512016 *Dipartimento Universitario*

**BOZZA nuova Copertina modello informazione rischi 1.2 CORRETTA SESSIMI**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001032790-untscl-8dc5cc64-9ce3-4614-8fbf-6d339055de9a-003.pdf  
Salute Via Monte Cengio (Palestra C.U.S.) Via Manzoni 16 - *Dipartimento Universitario Clinico* di

**001032229-untscl-e64969b4-ca2e-4812-85ac-97a3fe12c709-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001032229-untscl-e64969b4-ca2e-4812-85ac-97a3fe12c709-000.pdf  
del Consiglio del *Dipartimento Universitario Clinico* di Scienze mediche, chirurgiche e della salute

**01 DOMANDA DI ACCREDITAMENTO FOTOGRAFI 2016 - IMPRESE**

**Frase chiave**  
corso di validità degli studi di  
del del d.p.r di un documento  
il dipartimento di in caso di  
in corso di legge studi di trieste

**Lingua**  
[Italian](#) (30)

Lingua dei documenti

# UI Solritas (III)

- Oltre alla mera ricerca, molte possibilità:

Solr browse: files - Mozilla Firefox

localhost:8983/solr/files/browse?q=dipartimento+universitario+clinico&type=pdf&fq=language%3A"it"

Cerca

Lingua di navigazione utente →

Cerca: dipartimento universitario clinico

> language:"it" [x](#)

30 risultati trovati in 44ms Pagina 1 di 3

Tutti i formati (30) PDF (30) File testuali (0)

**001026124-untscl-7bee8dc3-c8d9-49dc-b7ca-1c3e682cebf3-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001026124-untscl-7bee8dc3-c8d9-49dc-b7ca-1c3e682cebf3-000.pdf  
.201512016 *Dipartimento Universitario Clinico* di Scienze Mediche, Ghirurgiche e della Salute Si rende

**001018758-untscl-a4051605-0159-4471-9fc6-1a03b52fadfb-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001018758-untscl-a4051605-0159-4471-9fc6-1a03b52fadfb-000.pdf  
PREMIO DI STUDIO IN MEMORIA DI LAURA VITTORI Anno Accademico 201 512016 *Dipartimento Universitario*

**BOZZA nuova Copertina modello informazione rischi 1.2 CORRETTA SESSIMI**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001032790-untscl-8dc5cc64-9ce3-4614-8fbf-6d339055de9a-003.pdf  
Salute Via Monte Cengio (Palestra C.U.S.) Via Manzoni 16 - *Dipartimento Universitario Clinico* di

**001032229-untscl-e64969b4-ca2e-4812-85ac-97a3fe12c709-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001032229-untscl-e64969b4-ca2e-4812-85ac-97a3fe12c709-000.pdf  
del Consiglio del *Dipartimento Universitario Clinico* di Scienze mediche, chirurgiche e della salute

**01 DOMANDA DI ACCREDITAMENTO FOTOGRAFI 2016 - IMPRESE**

Frase chiave

corso di validità degli studi di  
del del d.p.r di un documento  
il dipartimento di in caso di  
in corso di legge studi di trieste

Lingua  
[Italian](#) (30)

Lingua dei documenti



# UI Solritas (III)

- Oltre alla mera ricerca, molte possibilità:

Solr browse: files - Mozilla Firefox

localhost:8983/solr/files/browse?q=dipartimento+universitario+clinico&type=pdf&fq=language%3A"it"

Cerca

Lingua di navigazione utente →

Cerca: dipartimento universitario clinico

> language:"it" [x](#)

Tipi di file

30 risultati trovati in 44ms Pagina 1 di 3

Tutti i formati (30) PDF (30) File testuali (0)

**001026124-untscl-7bee8dc3-c8d9-49dc-b7ca-1c3e682cebf3-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001026124-untscl-7bee8dc3-c8d9-49dc-b7ca-1c3e682cebf3-000.pdf  
.201512016 *Dipartimento Universitario Clinico* di Scienze Mediche, Ghirurgiche e della Salute Si rende

**001018758-untscl-a4051605-0159-4471-9fc6-1a03b52fadfb-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001018758-untscl-a4051605-0159-4471-9fc6-1a03b52fadfb-000.pdf  
PREMIO DI STUDIO IN MEMORIA DI LAURA VITTORI Anno Accademico 201 512016 *Dipartimento Universitario*

**BOZZA nuova Copertina modello informazione rischi 1.2 CORRETTA SESSIMI**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001032790-untscl-8dc5cc64-9ce3-4614-8fbf-6d339055de9a-003.pdf  
Salute Via Monte Cengio (Palestra C.U.S.) Via Manzoni 16 - *Dipartimento Universitario Clinico* di

**001032229-untscl-e64969b4-ca2e-4812-85ac-97a3fe12c709-000.pdf**  
id: /media/sf\_VirtualBoxShared/DocumentiCondivisiSolr/001032229-untscl-e64969b4-ca2e-4812-85ac-97a3fe12c709-000.pdf  
del Consiglio del *Dipartimento Universitario Clinico* di Scienze mediche, chirurgiche e della salute

**01 DOMANDA DI ACCREDITAMENTO FOTOGRAFI 2016 - IMPRESE**

Frase chiave

corso di validità degli studi di  
del del d.p.r di un documento  
il dipartimento di in caso di  
in corso di legge studi di trieste

Lingua

[Italian](#) (30)

Lingua dei documenti

# Riconoscimento PDF già indicizzabili (I)

- La presenza della sottostringa „Font“ è il discriminante

```
endobj
10 0 obj
<</BaseFont/CFDNCX+Helvetica-Narrow/FontDescriptor 11 0 R/Type/Font
/FirstChar 32/LastChar 116/Widths[
228 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 228 0
456 456 456 456 456 0 0 0 456 456 228 0 0 0 0
0 547 0 592 0 0 501 0 0 228 0 0 0 0 592 0
547 0 0 0 0 0 547 0 0 0 0 0 0 0 0 0
0 456 0 410 456 456 0 0 0 182 0 0 182 0 0 456
0 0 273 410 228]
/Encoding/WinAnsiEncoding/Subtype/Type1>>
endobj
8 0 obj
<</BaseFont/PLITTA+Helvetica/FontDescriptor 9 0 R/Type/Font
/FirstChar 32/LastChar 117/Widths[
278 0 0 0 0 0 0 0 0 0 0 0 0 0 0 278 0
556 556 556 556 556 556 556 556 556 556 0 0 0 0 0
0 667 667 722 722 667 611 778 722 278 500 0 556 833 722 778
667 778 722 667 611 722 667 0 0 0 611 0 0 0 0
0 556 0 0 0 556 0 556 0 222 0 0 222 0 556 556
0 0 0 0 278 556]
/Encoding/WinAnsiEncoding/Subtype/Type1>>
endobj
11 0 obj
<</Type/FontDescriptor/FontName/CFDNCX+Helvetica-Narrow/FontBBox[0 -19 558 737]/Flags 131106
/Ascent 737
/CapHeight 737
/Descent -19
/ItalicAngle 0
/StemV 86
/MissingWidth 228
/XHeight 538
/CharSet(/A/C/F/I/N/P/V/a/c/colon/d/e/eight/four/i/l/nine/o/one/period/r/s/space/t/three/two/zero)/FontFile3 16 0 R>>
endobj
16 0 obj
<</Filter/FlateDecode
/Subtype/Type1C/Length 2023>>stream
xe•PSW•Çiã%/OQo•N½•[•Q-ø•••••E•••••%•••••@~ ÚUTô;ç@¿©ÁÊ/A•-V•••••?©»ëihm•ûóø«;ûóéìì?÷p9çÛ¹ç|iç••••\•A•Ã•iöYÚL]çzr,Úh4d;P½'ÁváÇ\•ßÖ¿
¥••••xë••••Éúûö•
```



## Riconoscimento PDF già indicizzabili (II)

```
La stringa 'Font' e' stata trovata 0 volte  
Il file /media/sf_VirtualBoxShared/DocumentiCondivisiSolr/000018778-untscle-  
d47687b1-78fd-493f-9741-a594295314c5-000.pdf NON e' ancora indicizzabile
```

```
La stringa 'Font' e' stata trovata 8 volte  
Il file /media/sf_VirtualBoxShared/DocumentiCondivisiSolr/000987779-untscle-  
29734f51-cccd-4d7e-adf4-f502dbec02c6-000.pdf e' gia' indicizzabile
```

## Riconoscimento PDF già indicizzabili (II)

0



```
La stringa 'Font' e' stata trovata 0 volte  
Il file /media/sf_VirtualBoxShared/DocumentiCondivisiSolr/000018778-untscle-  
d47687b1-78fd-493f-9741-a594295314c5-000.pdf NON e' ancora indicizzabile
```

```
La stringa 'Font' e' stata trovata 8 volte  
Il file /media/sf_VirtualBoxShared/DocumentiCondivisiSolr/000987779-untscle-  
29734f51-cccd-4d7e-adf4-f502dbec02c6-000.pdf e' gia' indicizzabile
```

## Riconoscimento PDF già indicizzabili (II)

```
La stringa 'Font' e' stata trovata 0 volte  
Il file /media/sf_VirtualBoxShared/DocumentiCondivisiSolr/000018778-untscle-  
d47687b1-78fd-493f-9741-a594295314c5-000.pdf NON e' ancora indicizzabile
```

>0



```
La stringa 'Font' e' stata trovata 8 volte  
Il file /media/sf_VirtualBoxShared/DocumentiCondivisiSolr/000987779-untscle-  
29734f51-cccd-4d7e-adf4-f502dbec02c6-000.pdf e' gia' indicizzabile
```

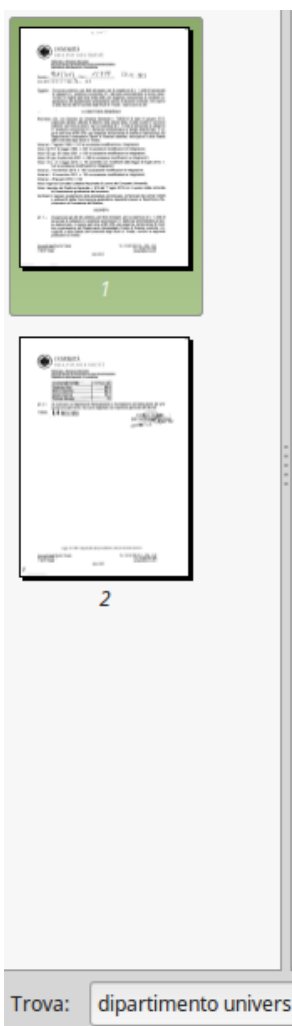
## Riconoscimento PDF già indicizzabili (II)

```
La stringa 'Font' e' stata trovata 0 volte  
Il file /media/sf VirtualBoxShared/DocumentiCondivisiSolr/000018778-untscle-  
d47687b1-78fd-493f-9741-a594295314c5-000.pdf NON e' ancora indicizzabile
```

- Il file pdf non ancora indicizzabile viene indicizzato su Solr solo dopo esser stato reso indicizzabile tramite OCR

# OCR (*Optical Character Recognition*)

- Consente di estrarre il testo da un file scansionato
- Rendendolo indicizzabile per contenuto



**UNIVERSITÀ  
DEGLI STUDI DI TRIESTE**

**Rettorato e Direzione Generale  
Sezione Servizi al Personale Tecnico-Amministrativo  
Ripartizione Reclutamento e Formazione**

Decreto n. 928/2013 - Prot. n. 157 PP 08 AGO 2013  
Anno 2013 tit. VII cl. 1 fasc. 5 All. 0

**Oggetto:** Concorso pubblico, per titoli ed esami, per la copertura di n. 1 unità di personale di categoria C, posizione economica C1, dell'area amministrativa a tempo determinato, in regime part-time all'83,33%, per esigenze, temporanee di carattere organizzativo del **Dipartimento Universitario Clinico** di Scienze mediche, chirurgiche e della Salute dell'Università degli Studi di Trieste - Approvazione atti

**IL DIRETTORE GENERALE**

Promosso che, con Decreto del Direttore Generale n. 739/2013 di data 14 giugno 2013,

Trova:  < Trova precedente > Trova successivo Maiuscole/minuscole 3 corrispondenze in questa pagina

# OCR (Optical Character Recognition)

Tesseract Open Source OCR Engine v3.04.01 with Leptonica

Page 0 : /tmp/000963750-untscle-cade750d-584c-4d45-8b8e-79491187e6b3-000\_OUTPUT-0.jpg

Page 1 : /tmp/000963750-untscle-cade750d-584c-4d45-8b8e-79491187e6b3-000\_OUTPUT-1.jpg

Page 2 : /tmp/000963750-untscle-cade750d-584c-4d45-8b8e-79491187e6b3-000\_OUTPUT-2.jpg

Page 3 : /tmp/000963750-untscle-cade750d-584c-4d45-8b8e-79491187e6b3-000\_OUTPUT-3.jpg

Page 4 : /tmp/000963750-untscle-cade750d-584c-4d45-8b8e-79491187e6b3-000\_OUTPUT-4.jpg

Page 5 : /tmp/000963750-untscle-cade750d-584c-4d45-8b8e-79491187e6b3-000\_OUTPUT-5.jpg

Page 6 : /tmp/000963750-untscle-cade750d-584c-4d45-8b8e-79491187e6b3-000\_OUTPUT-6.jpg



UNIVERSITÀ  
DEGLI STUDI DI TRIESTE

Rettorato e Direzione Generale  
Sezione Servizi al Personale Tecnico-Amministrativo  
Ripartizione Reclutamento e Formazione

Decreto n. 928/2013 - Prot. n. 157 PP  
Anno 2013 tit. VII cl. 1 fasc. 5 All. 0

08 AGO 2013

Oggetto: Concorso pubblico, per titoli ed esami, per la copertura di n. 1 unità di personale di categoria C, posizione economica C1, dell'area amministrativa a tempo determinato, in regime part-time all'83,33%, per esigenze, temporanee di carattere organizzativo del **Dipartimento Universitario Clinico** di Scienze mediche, chirurgiche e della Salute dell'Università degli Studi di Trieste - Approvazione atti

IL DIRETTORE GENERALE

Promosso con Decreto del Direttore Generale n. 739/2013 di data 14 giugno 2013



## Sviluppi futuri

- Possibilità di ricercare tra intervallo di date, esiti votazione, categorie...
- Generalizzare l'ambiente
  - Non più virtual machine ma server „fisico“
- Poter aprire i documenti direttamente dalla UI Solritas
- Migliorare la grafica della UI Solritas
- Sperimentare con migliaia di file
- ...

# Grazie per l'attenzione



UNIVERSITÀ  
DEGLI STUDI DI TRIESTE

Cerca: