A Search Engine For The IVOA Documentation

- **This project has been achieved in a student internship.**
  - Sinan Acar (UTBM Belfort France) supervised by Laurent Michel
- http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

Working with the VO Documentation

- **46 standards + 50 notes**
  - Some in multiple versions
  - Number of dependencies increases over the time
  - Some not really used
- **Difficult to get a clear overview on the standards**
  - Personal opinion ...

http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

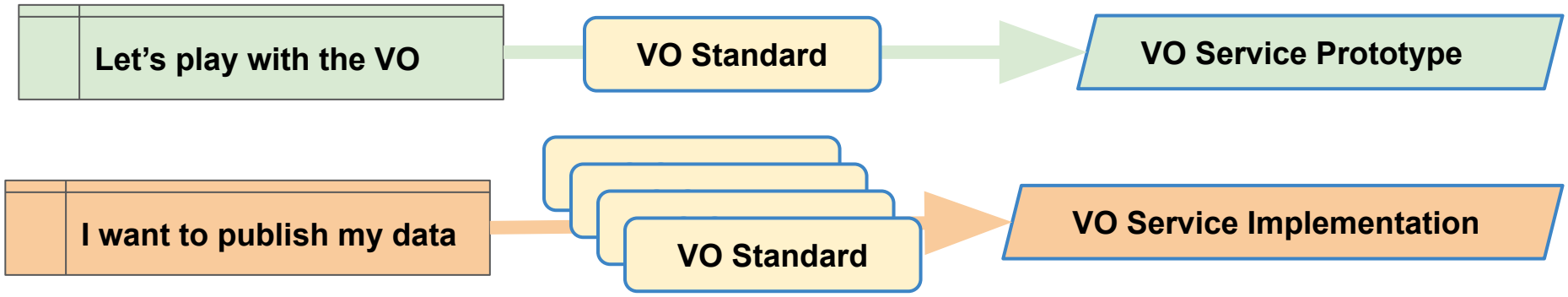
Working with the VO Documentation

- **46 standards + 60 notes**
  - Some in multiple versions
  - Number of dependencies increases over the time
  - Some not really used
- **Difficult to get a clear overview on the standards**

http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

Working with the VO Documentation

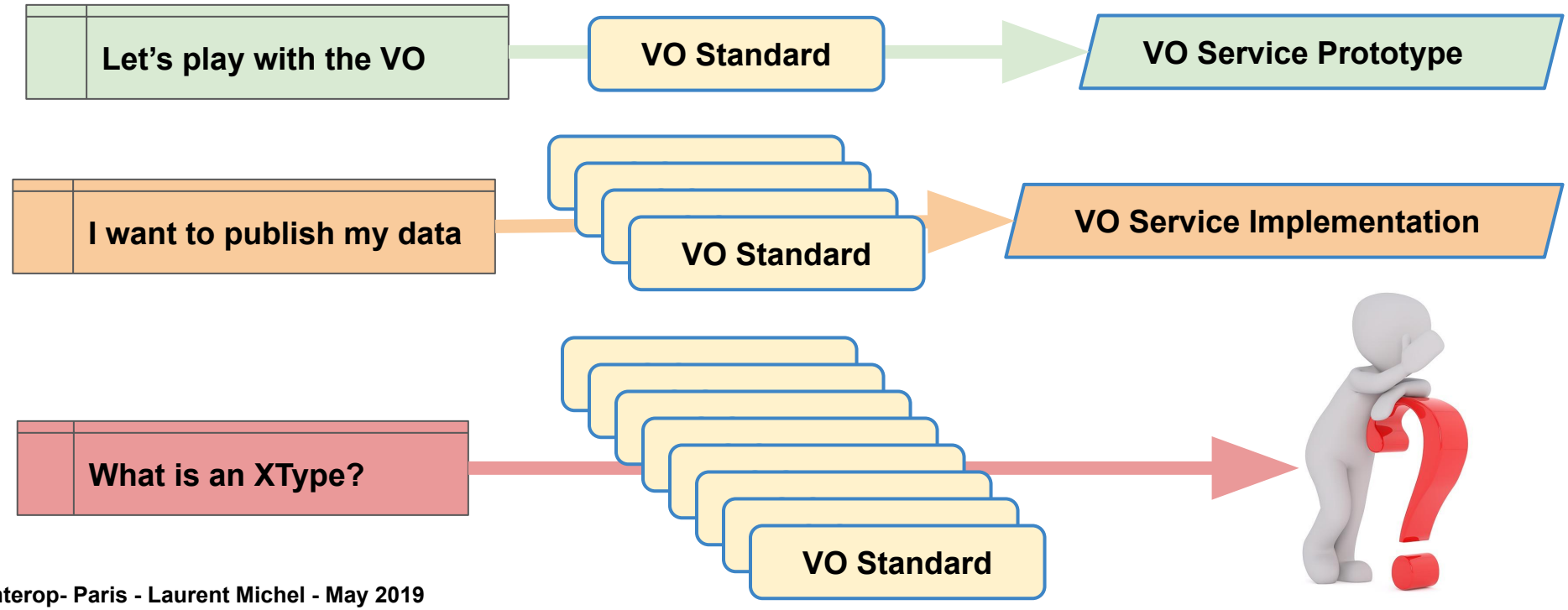
- **46 standards + 60 notes**
  - Some in multiple versions
  - Number of dependencies increases over the time
  - Some not really used
- **Difficult to get a clear overview on the standards**



http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

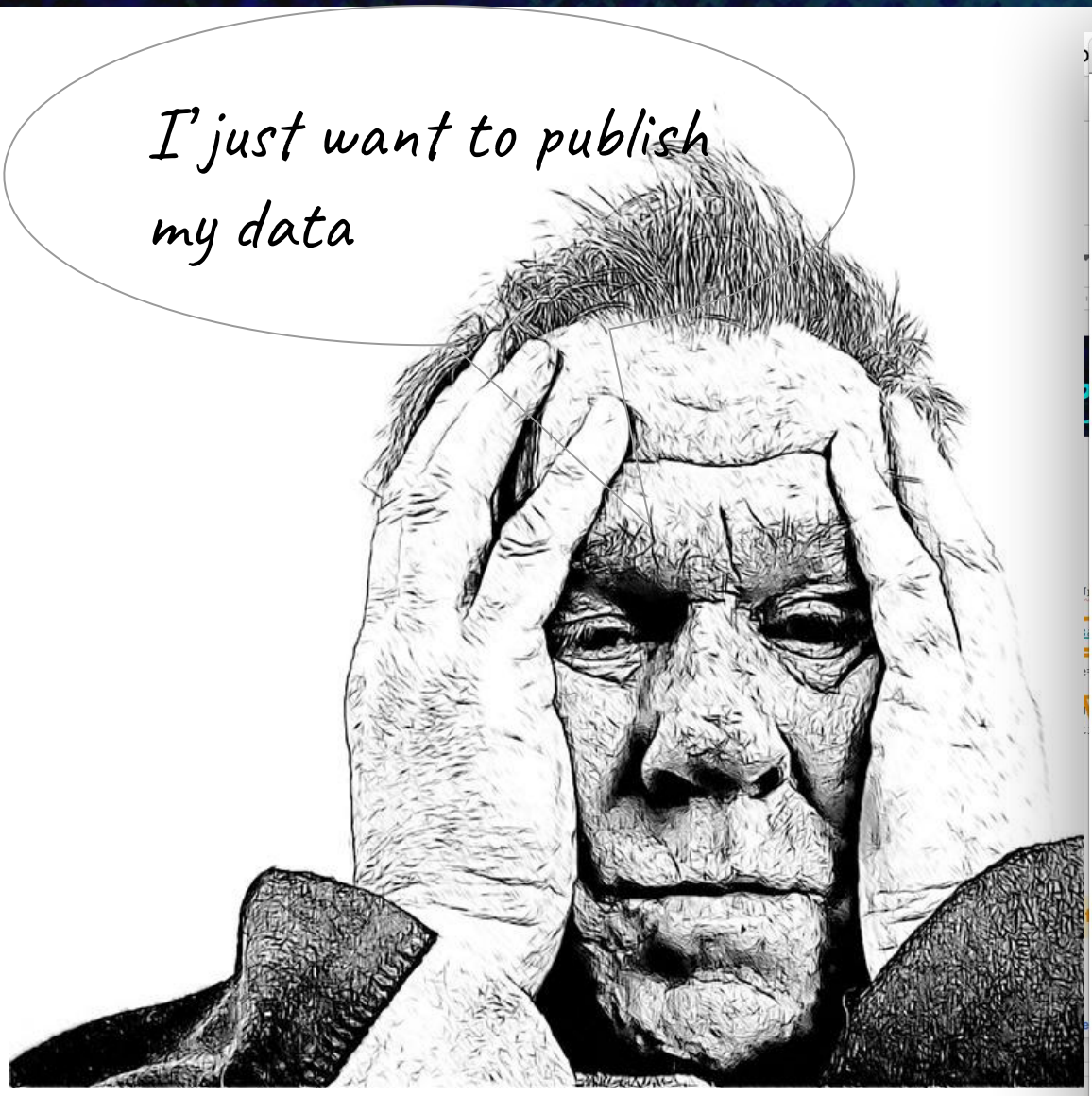
Working with the VO Documentation

- **46 standards + 60 notes**
  - Some in multiple versions
  - Number of dependencies increases over the time
  - Some not really used
- **Difficult to get a clear overview on the standards**



http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

If I Were in Bad Faith (for the fun)



laurentmichel — michel@galhecos:/rawdata/doc...

```
REC-DALI-1.0-20131129.pdf
REC-DALI-1.1.pdf
REC-DataLink-1.0-20150617.pdf
REC-DocStd-1.0-20031024.pdf
REC-DocStd-1.2.pdf
REC-DocStd-2.0-20170517.pdf
REC-HIPS-1.0-20170519.pdf
REC-Identifiers-2.0.pdf
REC-MOC-1.0-20140602.pdf
REC-ObsCore-v1.0-20111028.pdf
REC-ObsCore-v1.1-20170509.pdf
REC-PDL-1.0-20140523.pdf
REC-PhotDM-1.0-20131005.pdf
REC-RegistryInterface-1.0.pdf
REC-RegistryInterface-1.1.pdf
REC-RegTAP-1.0.pdf
REC-RM-1.01-20040426.pdf
REC-RM-1.12-20070302.pdf
REC-SAMP-1.2-20101216.pdf
REC-SAMP-1.3-20120411.pdf
REC-SIA-1.0.pdf
REC-SIA-2.0-20151223.pdf
REC-SimDAL-1.0-20170320.pdf
REC-SimpleDALRegExt-1.1.pdf
REC-SimpleDALRegExt-20131005.pdf
REC-SimulationDataModel-1.00-20120503.pdf
REC-SLAP-1.0-20101209.pdf
REC-SODA-1.0.pdf
REC-SpectrumDM-1.1-20111120.pdf
REC-SSA-1.1-20120210.pdf
REC-SSLDM-1.0-20101202.pdf
REC-SSOAuthMech-2.0.pdf
REC-StandardsRegExt-1.0-20120508.pdf
REC-TAP-1.0.pdf
REC-TAPRegExt-1.0.pdf
REC-UCDlist-1.3-20180527.pdf
REC-UWS-1.0-20101010.pdf
REC-UWS-1.1-20161024.pdf
REC-VODataService-1.1-20101202.pdf
REC-VOEvent-2.0.pdf
REC-VOResource-1.1.pdf
REC-VOSI-1.0-20110531.pdf
REC-VOSI-1.1.pdf
REC-VOSpace-1.15.pdf
REC-VOSpace-2.0-20130329.pdf
REC-VOSpace-2.1.pdf
REC-VOTable-1.2.pdf
```

http://saada.unistra.fr/esdoc/interface

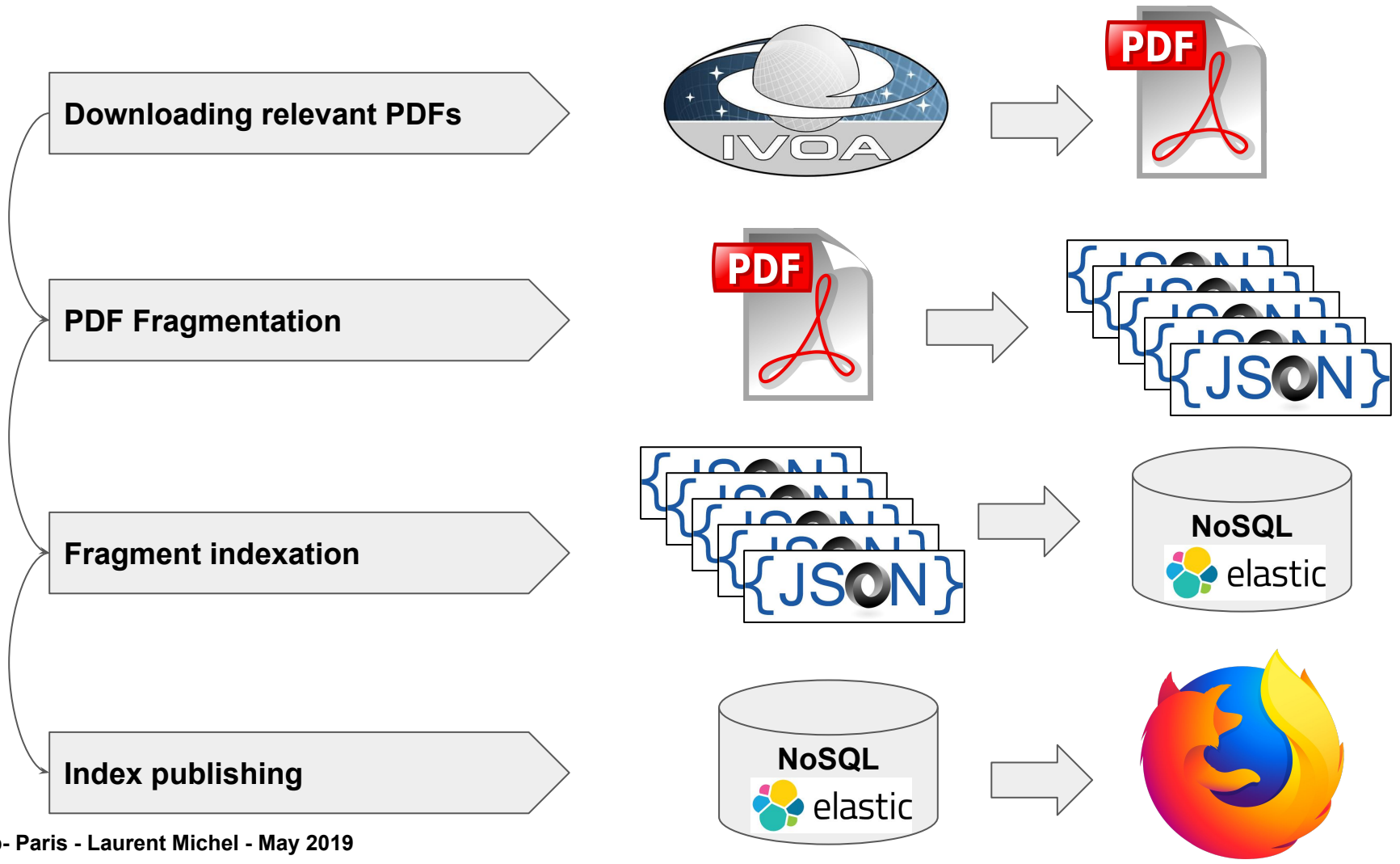
Being Objective

- **People can get much help from an helpful community**
  - Architecture document
  - `How to publish` page
  - Tools and frameworks
  - WG lists

- **I propose a little extra facility…**
  - **Finding text locations possibly answering my questions**
    - What is an XType?
    - Is the TAP_SCHEMA mandatory?

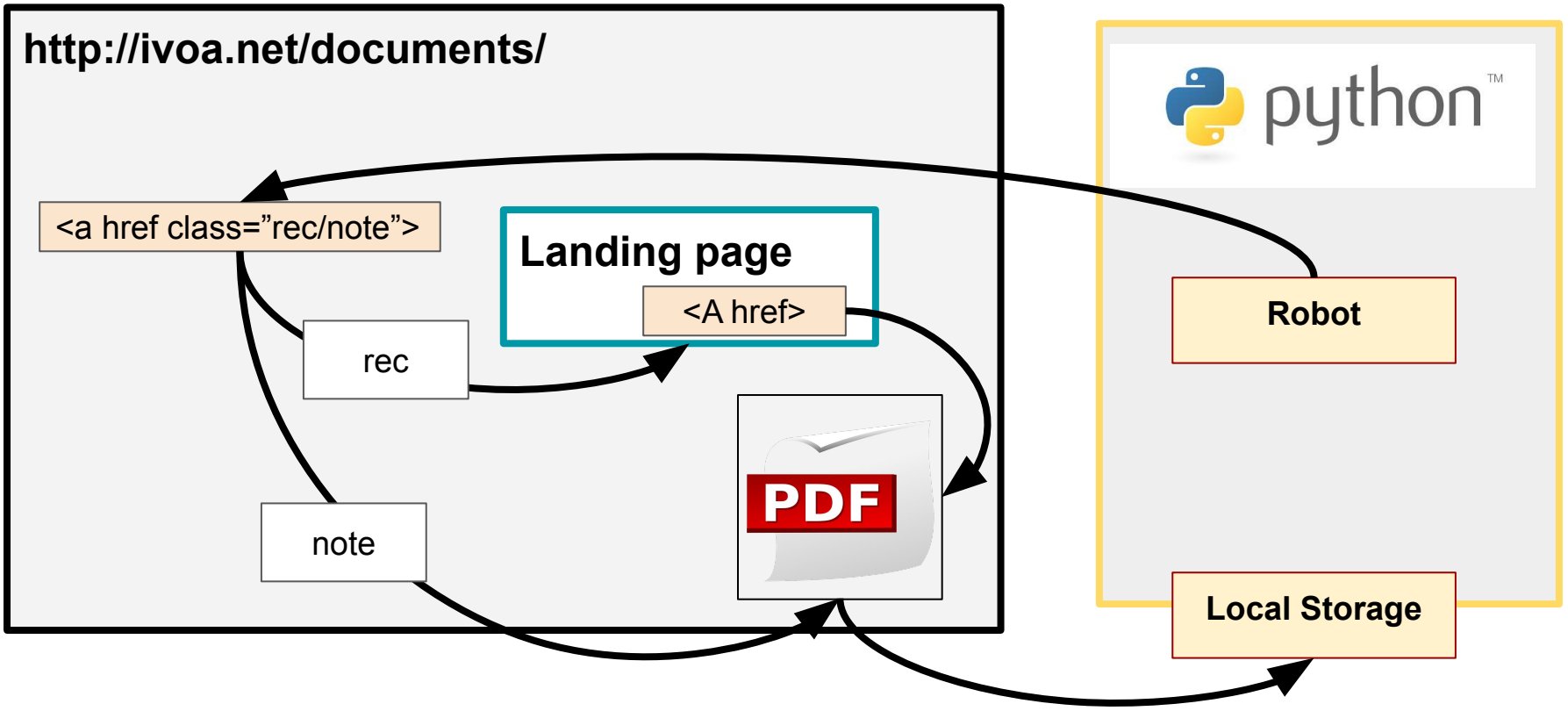
http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

A 4 Steps Process



http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

Step 1: Downloading PDFs



- **Based on CSS selectors**
  - **from** pyquery **import** PyQuery **as** pq
  - **$(".rec") ->** pointers to recommendations landing pages
  - **$(".note") ->** pointers to notes (no landing page)

http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

Step 2: PDF Fragmentation

- **PDF is a printer format**
  - **Page layout independent from the display device**
    - Reliable page numbering
    - Reliable fragment locations in pages
  - **Text structure lost**
    - No easy way to make the difference between a caption and a title e.g.

- **Parsing PDFs**
  - **Using Grobid (P. Lopez)**
    - Java application modified for our purpose
  - **Machine learning algorithm**
    - Split the text in short fragments
    - Infer the nature of each fragment
      - section, title, caption...
    - Give the location of each fragment
      - Page number + coordinates within the page

http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

Step 2bis: Text Fragmentation: PDF

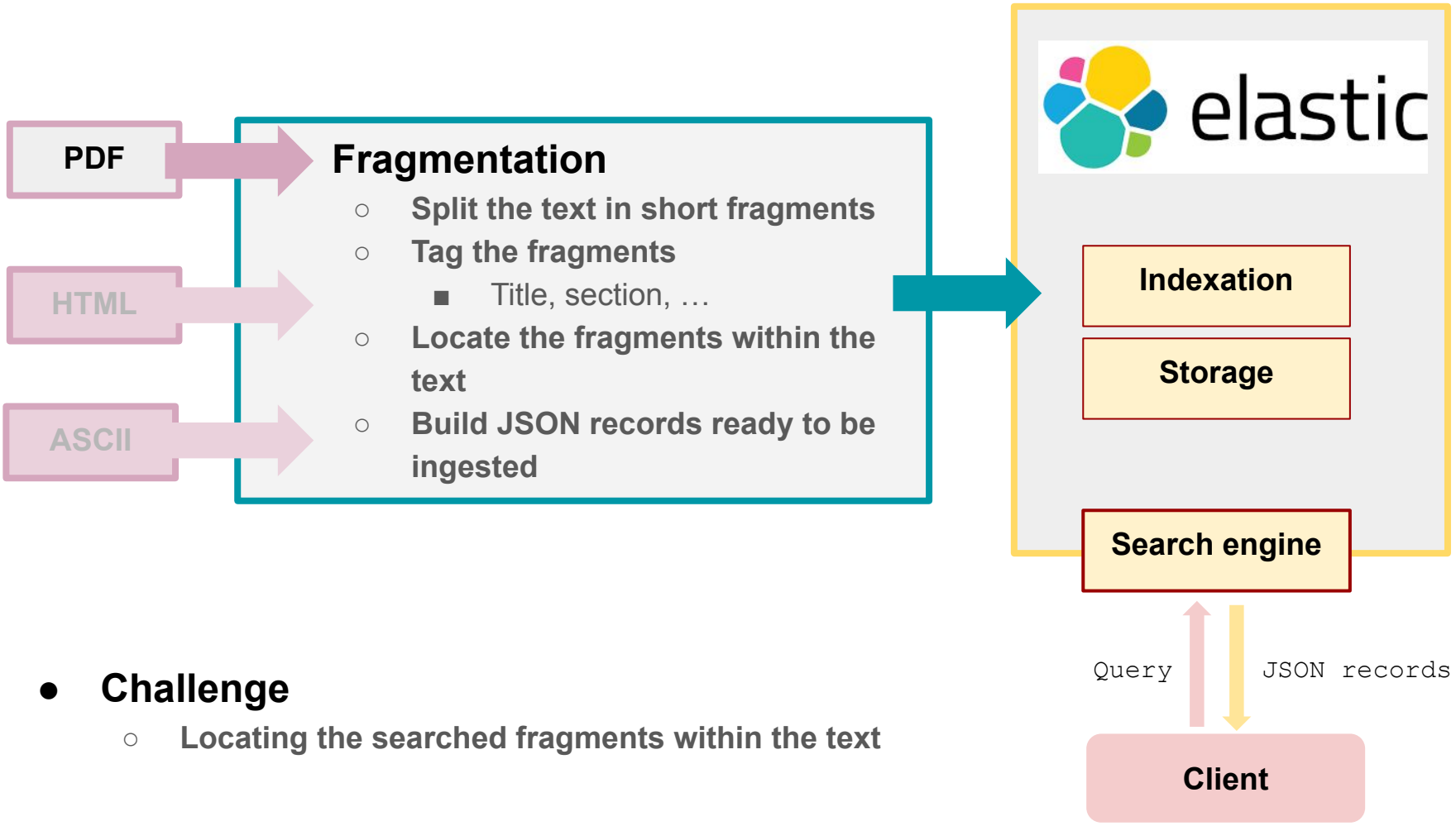

```
{
  "page": "2",
  "content": ". . . . . . . . . . . . . . . . . . . . . . . ",
  "filename": "VOUnits-REC-1.0-20140523.pdf"

  "url": "http://saada.unistra.fr/esdoc/corpus//IVOA/VOUnits-REC-1.0-20140523.pdf",

  "coords": "&page=2&y1=314.0",

  "title": "IVOA Units in the VO Version REC-1.0 IVOA",
  "element": "section",
  "type": "pdf"
}
```


Local copy of the document
- **No CORS issue**
- **No naming issue**

Fragment location within the document

http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

Step 3: Corpus Indexation

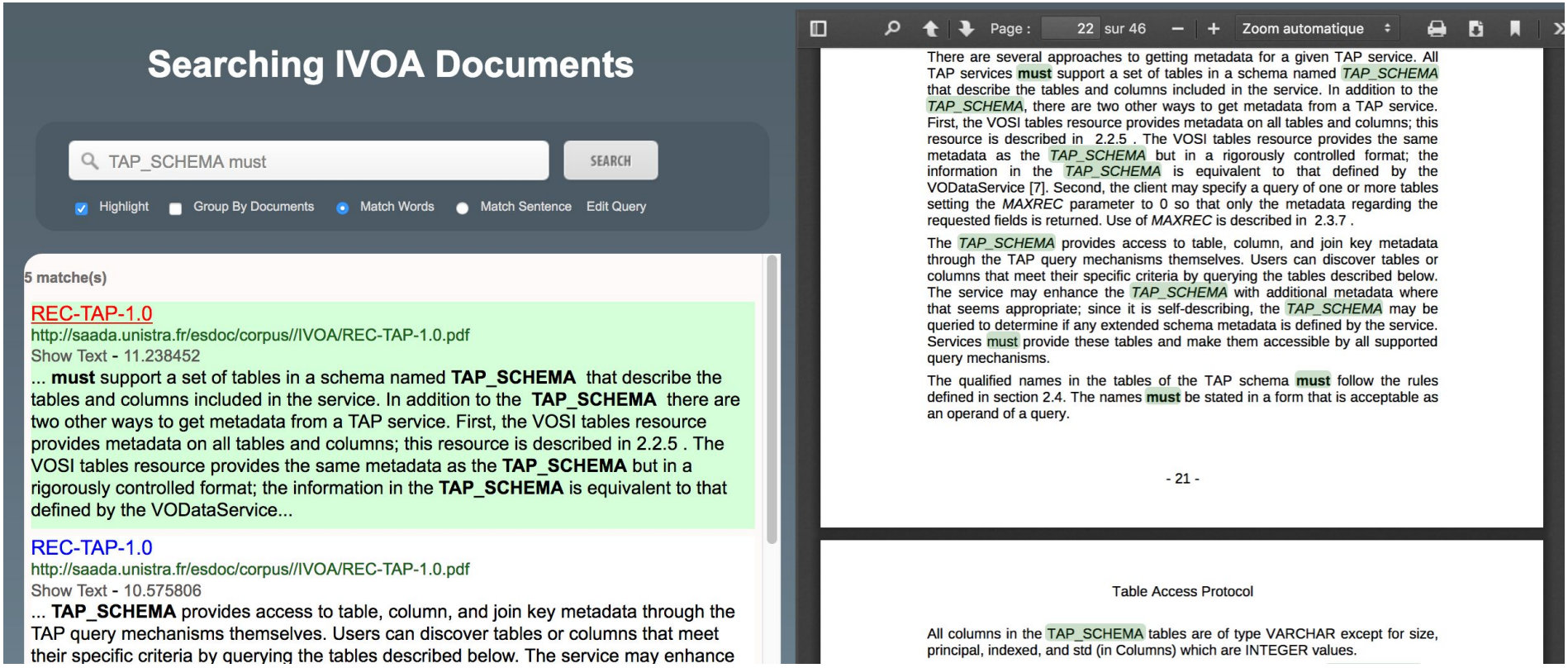


- **Challenge**
  - Locating the searched fragments within the text

http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

Step 4: Web Application

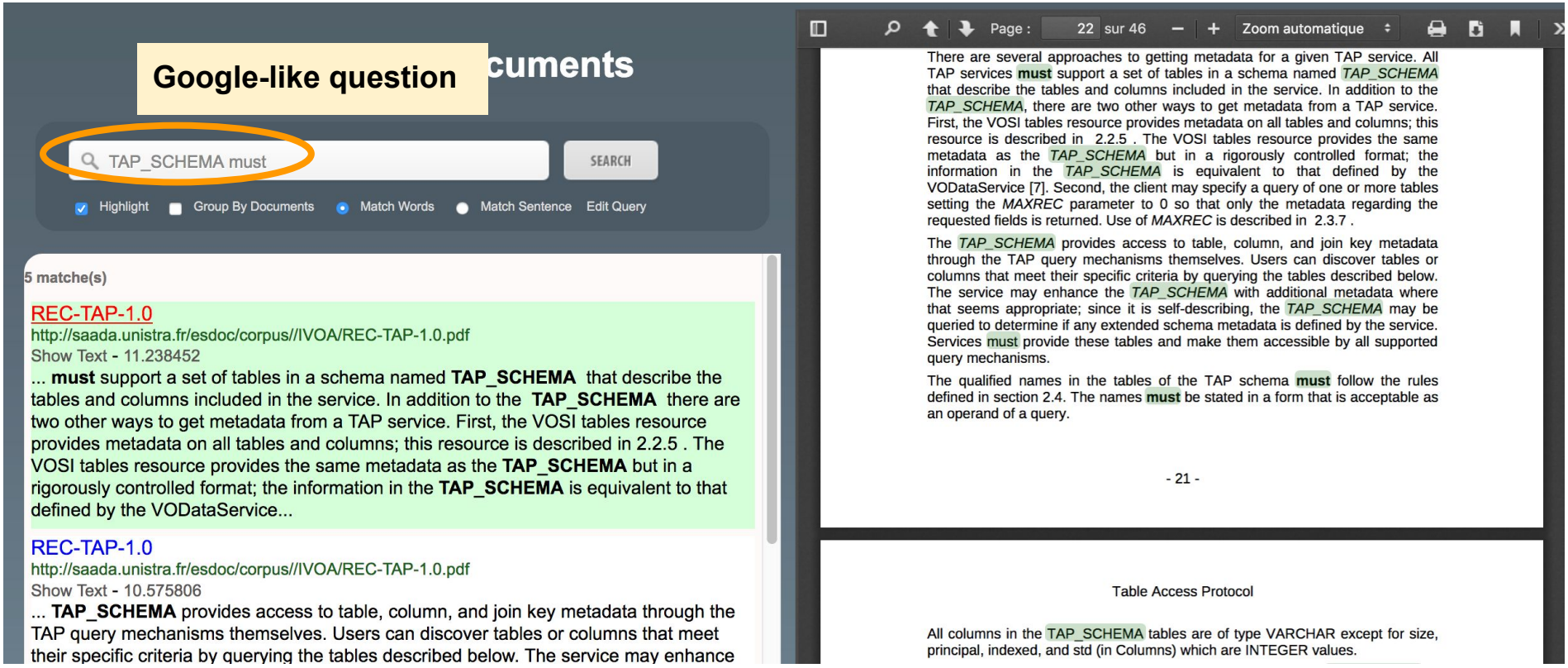
- **Question: Is the TAP_SCHEMA mandatory?**



http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

Step 4: Web Application

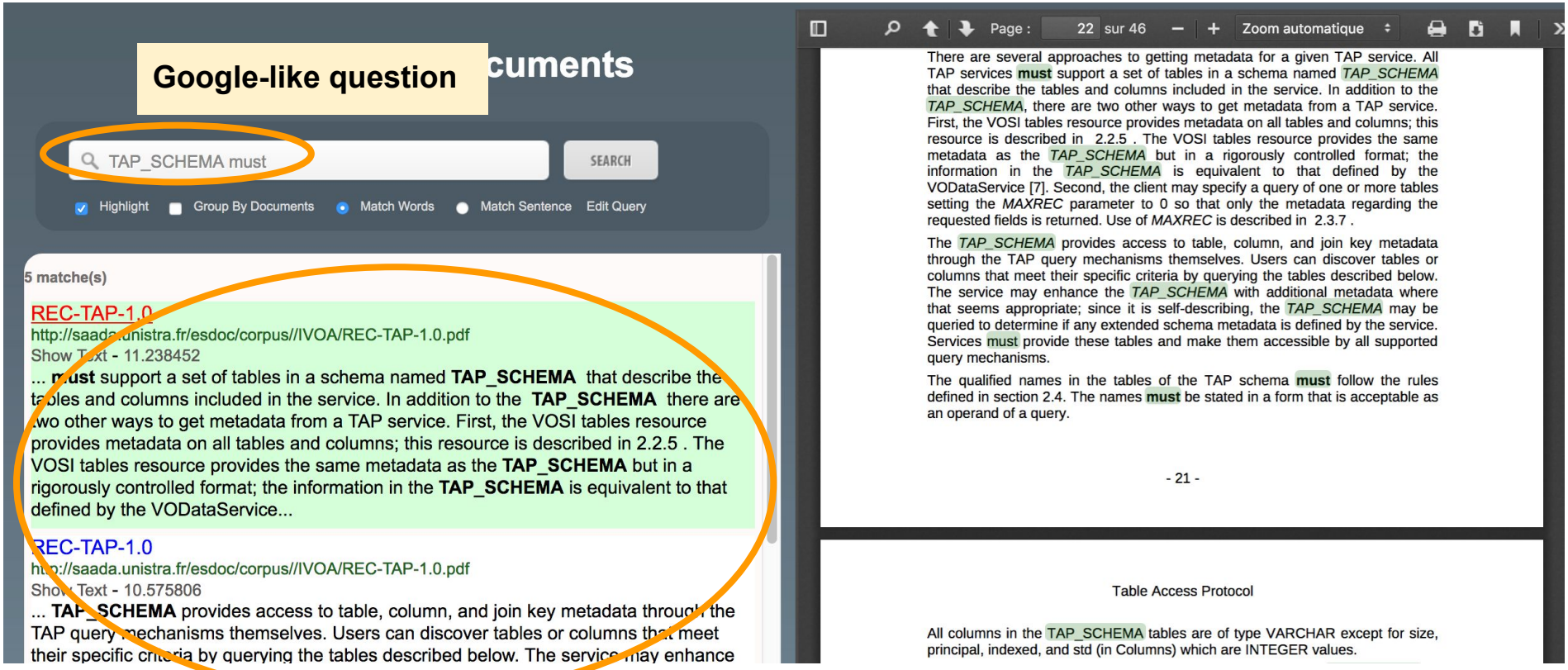
- **Question: Is the TAP_SCHEMA mandatory?**



http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

Step 4: Web Application

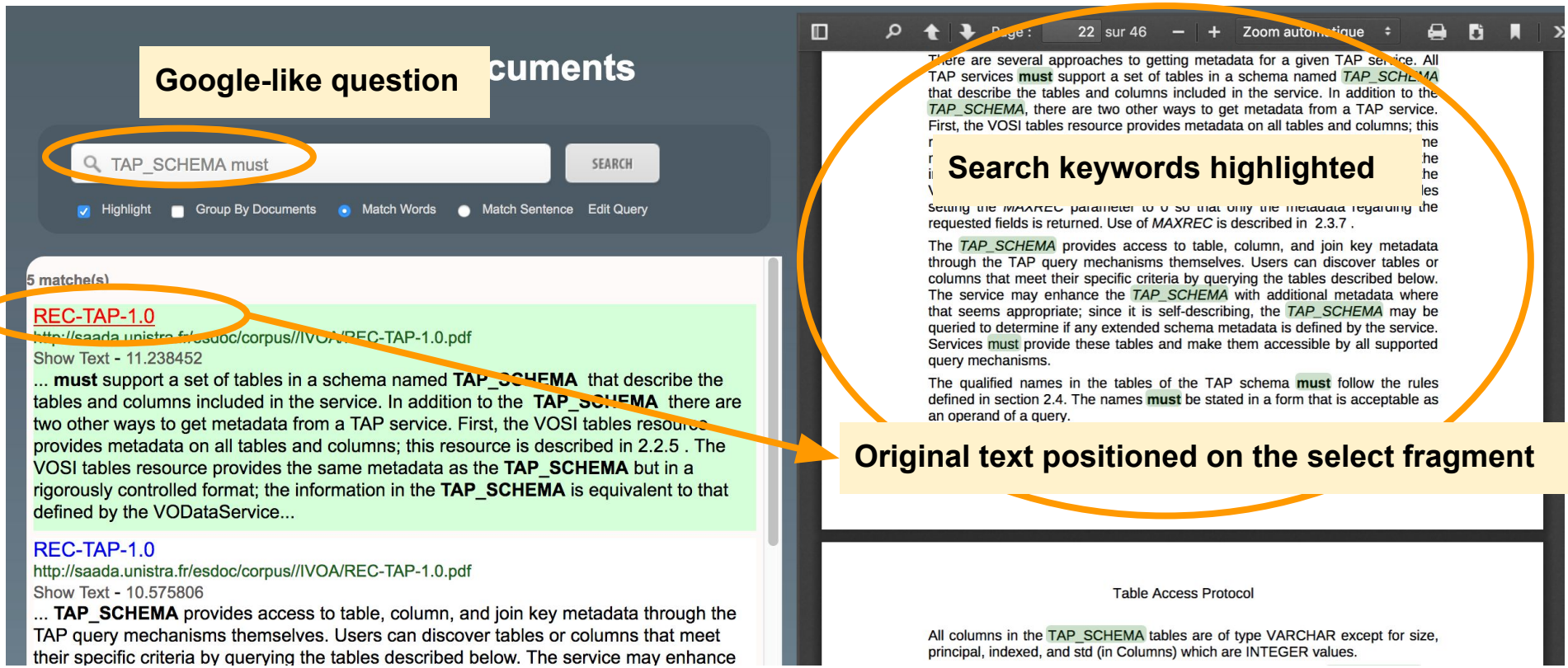
- **Question: Is the TAP_SCHEMA mandatory?**



http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

Step 4: Web Application

- **Question: Is the TAP_SCHEMA mandatory?**



http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

Thank You

http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa

http://saada.unistra.fr/esdoc/interfacePDF.html?index=ivoa