

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 ...

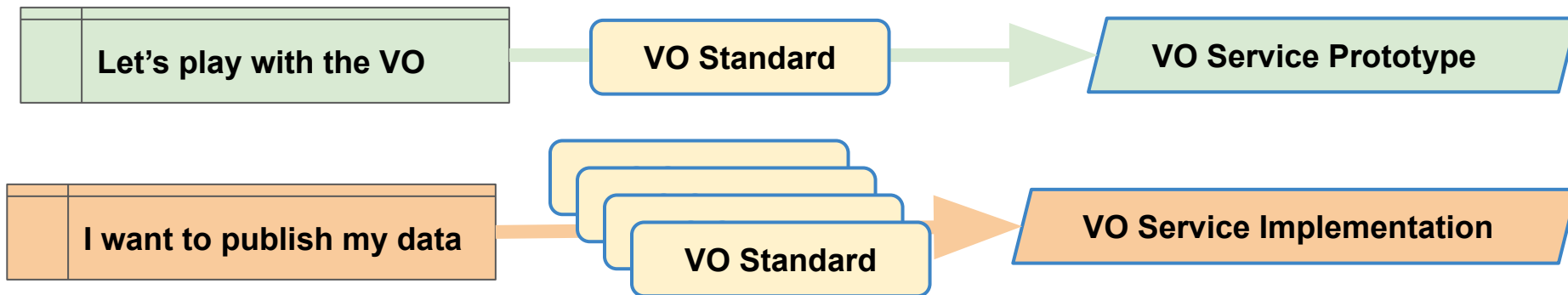
Working with the VO Documentation

- **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**



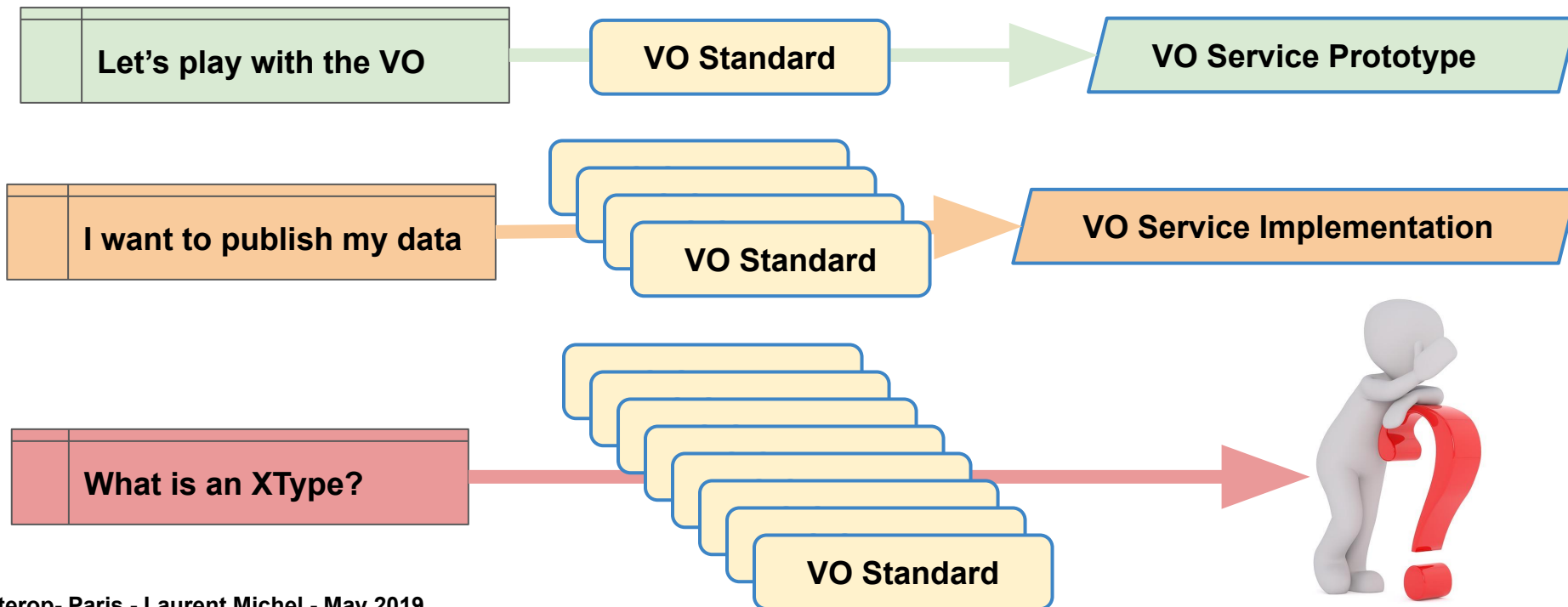
Working with the VO Documentation

- **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**



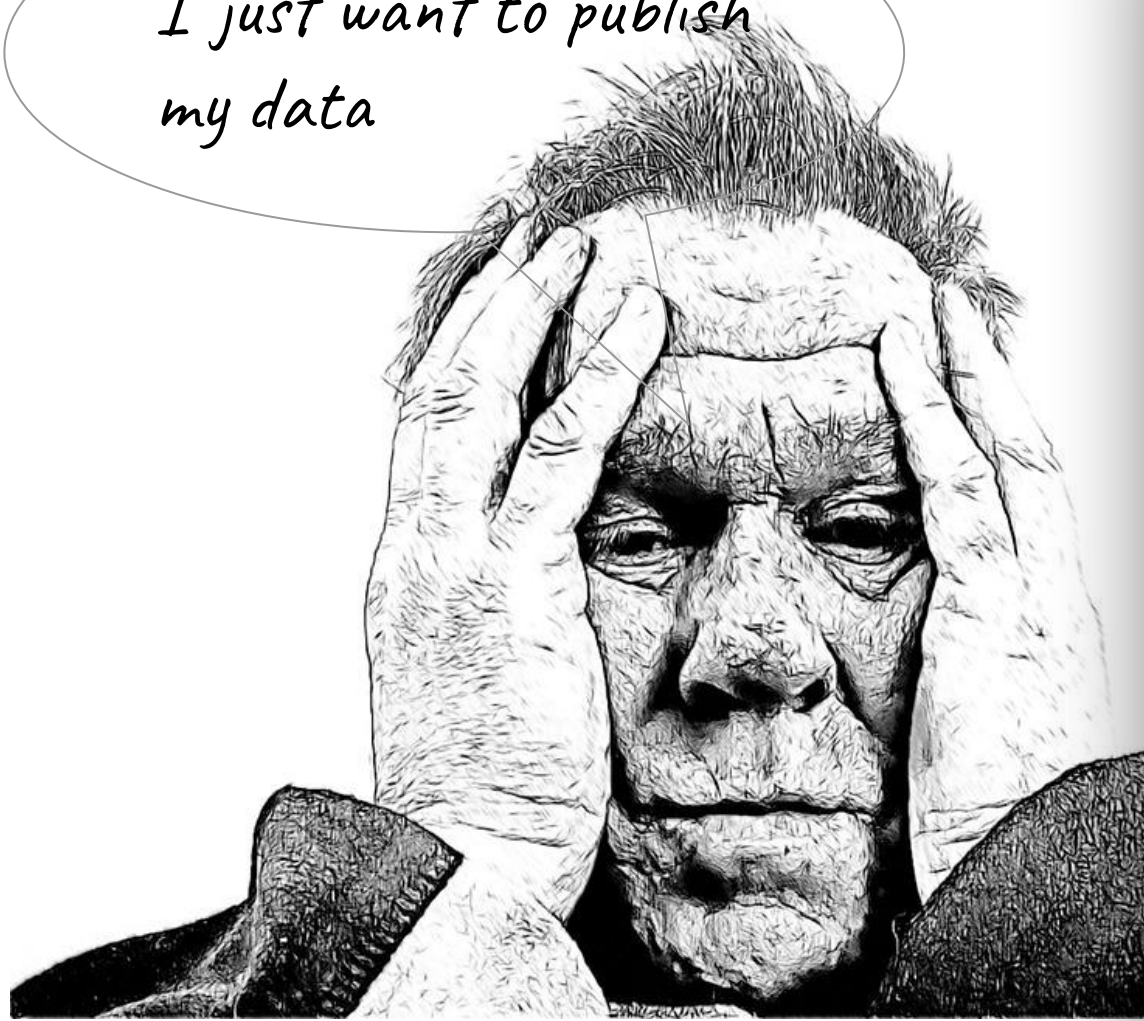
Working with the VO Documentation

- **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**



If I Were in Bad Faith (for the fun)

*I just want to publish
my data*



```
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-V0DataService-1.1-20101202.pdf
REC-V0Event-2.0.pdf
REC-V0Resource-1.1.pdf
REC-VOSI-1.0-20110531.pdf
REC-VOSI-1.1.pdf
REC-V0Space-1.15.pdf
REC-V0Space-2.0-20130329.pdf
REC-V0Space-2.1.pdf
REC-V0Table-1.2.pdf
```

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

Being Objective

- **People can get much help from an helpful community**

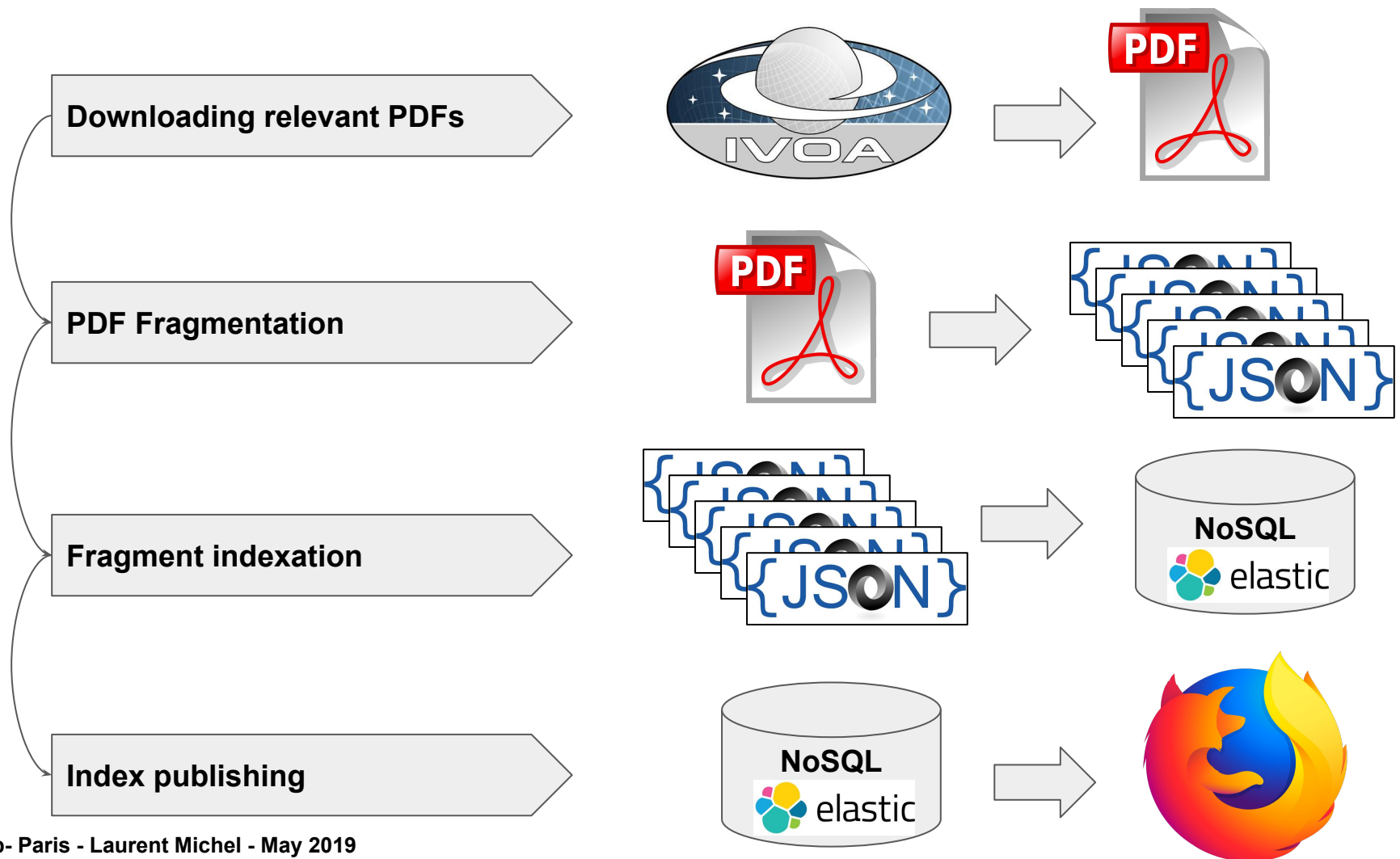
- Architecture document
- How to publish page
- Tools and frameworks
- WG lists



- **I propose a little extra facility...**

- **Finding text locations possibly answering my questions**
 - What is an XType?
 - Is the TAP_SCHEMA mandatory?

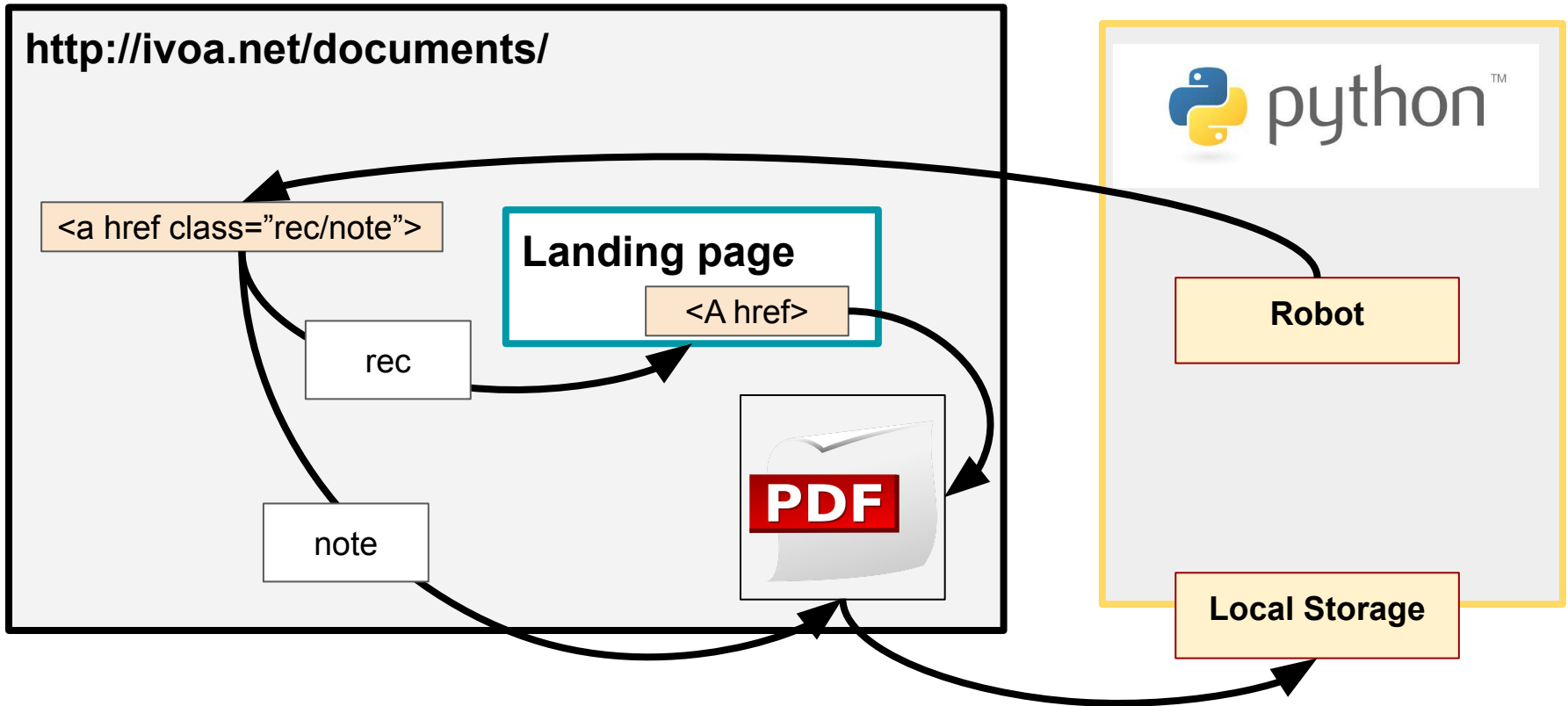
A 4 Steps Process



Interop- Paris - Laurent Michel - May 2019

<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)

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

Step 2bis: Text Fragmentation: PDF

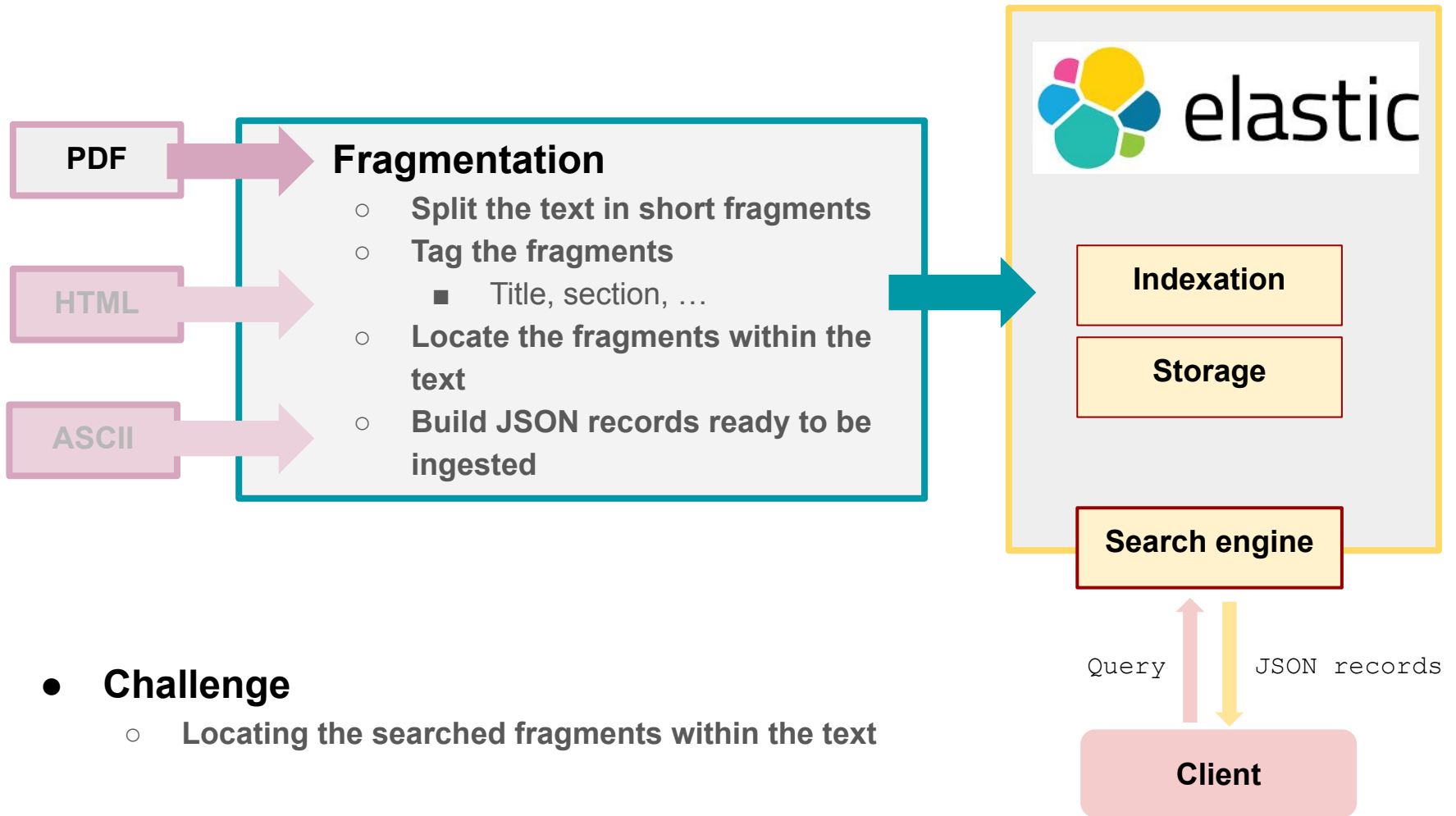
```
{  
  "page": "2",  
  "content": ". . . . .",  
  "filename": "VOUnits-REC-1.0-20140523.pdf"  
  
  "url": "http://saada.unistra.fr/esdoc/corpus//IVOA/VOUnits-REC-1.0-20140523.pdf",  
  
  "coords": "&page=2&y1=314.0",  
  
  "title": "IVOA Units in the VO Version REC-1.0 IVOA",  
  "element": "section",  
  "type": "pdf"  
}
```

Local copy of the document

- No CORS issue
- No naming issue

Fragment location within the document

Step 3: Corpus Indexation



Step 4: Web Application

- Question: Is the TAP_SCHEMA mandatory?

Searching IVOA Documents

Q TAP_SCHEMA must

SEARCH

Highlight Group By Documents Match Words Match Sentence [Edit Query](#)

5 matche(s)

REC-TAP-1.0

<http://saada.unistra.fr/esdoc/corpus//IVOA/REC-TAP-1.0.pdf>

Show Text - 11.238452

... **must** support a set of tables in a schema named **TAP_SCHEMA** that describe the tables and columns included in the service. In addition to the **TAP_SCHEMA** there are two other ways to get metadata from a TAP service. First, the VOSI tables resource provides metadata on all tables and columns; this resource is described in 2.2.5 . The VOSI tables resource provides the same metadata as the **TAP_SCHEMA** but in a rigorously controlled format; the information in the **TAP_SCHEMA** is equivalent to that defined by the VODataService...

REC-TAP-1.0

<http://saada.unistra.fr/esdoc/corpus//IVOA/REC-TAP-1.0.pdf>

Show Text - 10.575806

... **TAP_SCHEMA** provides access to table, column, and join key metadata through the TAP query mechanisms themselves. Users can discover tables or columns that meet their specific criteria by querying the tables described below. The service may enhance

Page : 22 sur 46 - + Zoom automatique

There are several approaches to getting metadata for a given TAP service. All TAP services **must** support a set of tables in a schema named **TAP_SCHEMA** that describe the tables and columns included in the service. In addition to the **TAP_SCHEMA**, there are two other ways to get metadata from a TAP service. First, the VOSI tables resource provides metadata on all tables and columns; this resource is described in 2.2.5 . The VOSI tables resource provides the same metadata as the **TAP_SCHEMA** but in a rigorously controlled format; the information in the **TAP_SCHEMA** is equivalent to that defined by the VODataService [7]. Second, the client may specify a query of one or more tables setting the **MAXREC** parameter to 0 so that only the metadata regarding the requested fields is returned. Use of **MAXREC** is described in 2.3.7 .

The **TAP_SCHEMA** provides access to table, column, and join key metadata through the TAP query mechanisms themselves. Users can discover tables or columns that meet their specific criteria by querying the tables described below. The service may enhance the **TAP_SCHEMA** with additional metadata where that seems appropriate; since it is self-describing, the **TAP_SCHEMA** may be queried to determine if any extended schema metadata is defined by the service. Services **must** provide these tables and make them accessible by all supported query mechanisms.

The qualified names in the tables of the TAP schema **must** follow the rules defined in section 2.4. The names **must** be stated in a form that is acceptable as an operand of a query.

- 21 -

Table Access Protocol

All columns in the **TAP_SCHEMA** tables are of type VARCHAR except for size, principal, indexed, and std (in Columns) which are INTEGER values.

Step 4: Web Application

- Question: Is the TAP_SCHEMA mandatory?

The image shows a search interface on the left and a PDF document on the right. The search interface has a search bar containing the text "TAP_SCHEMA must", which is circled in orange. Below the search bar are several filters: "Highlight" (checked), "Group By Documents" (unchecked), "Match Words" (checked), "Match Sentence" (unchecked), and "Edit Query". The search results show 5 matches, with the first two highlighted in green. The first match is titled "REC-TAP-1.0" and contains the text: "... **must** support a set of tables in a schema named **TAP_SCHEMA** that describe the tables and columns included in the service. In addition to the **TAP_SCHEMA** there are two other ways to get metadata from a TAP service. First, the VOSI tables resource provides metadata on all tables and columns; this resource is described in 2.2.5 . The VOSI tables resource provides the same metadata as the **TAP_SCHEMA** but in a rigorously controlled format; the information in the **TAP_SCHEMA** is equivalent to that defined by the VODataService...". The second match is also titled "REC-TAP-1.0" and contains the text: "... **TAP_SCHEMA** provides access to table, column, and join key metadata through the TAP query mechanisms themselves. Users can discover tables or columns that meet their specific criteria by querying the tables described below. The service may enhance".

The PDF document on the right is titled "Table Access Protocol" and contains the following text: "There are several approaches to getting metadata for a given TAP service. All TAP services **must** support a set of tables in a schema named **TAP_SCHEMA** that describe the tables and columns included in the service. In addition to the **TAP_SCHEMA**, there are two other ways to get metadata from a TAP service. First, the VOSI tables resource provides metadata on all tables and columns; this resource is described in 2.2.5 . The VOSI tables resource provides the same metadata as the **TAP_SCHEMA** but in a rigorously controlled format; the information in the **TAP_SCHEMA** is equivalent to that defined by the VODataService [7]. Second, the client may specify a query of one or more tables setting the **MAXREC** parameter to 0 so that only the metadata regarding the requested fields is returned. Use of **MAXREC** is described in 2.3.7 . The **TAP_SCHEMA** provides access to table, column, and join key metadata through the TAP query mechanisms themselves. Users can discover tables or columns that meet their specific criteria by querying the tables described below. The service may enhance the **TAP_SCHEMA** with additional metadata where that seems appropriate; since it is self-describing, the **TAP_SCHEMA** may be queried to determine if any extended schema metadata is defined by the service. Services **must** provide these tables and make them accessible by all supported query mechanisms. The qualified names in the tables of the TAP schema **must** follow the rules defined in section 2.4. The names **must** be stated in a form that is acceptable as an operand of a query."

Step 4: Web Application

- Question: Is the TAP_SCHEMA mandatory?

Google-like question documents

Q TAP_SCHEMA must SEARCH

Highlight Group By Documents Match Words Match Sentence Edit Query

5 matche(s)

REC-TAP-1.0
<http://saada.unistra.fr/esdoc/corpus//IVOA/REC-TAP-1.0.pdf>
Show Text - 11.238452

... **must** support a set of tables in a schema named **TAP_SCHEMA** that describe the tables and columns included in the service. In addition to the **TAP_SCHEMA** there are two other ways to get metadata from a TAP service. First, the VOSI tables resource provides metadata on all tables and columns; this resource is described in 2.2.5 . The VOSI tables resource provides the same metadata as the **TAP_SCHEMA** but in a rigorously controlled format; the information in the **TAP_SCHEMA** is equivalent to that defined by the VODataService...

REC-TAP-1.0
<http://saada.unistra.fr/esdoc/corpus//IVOA/REC-TAP-1.0.pdf>
Show Text - 10.575806

... **TAP_SCHEMA** provides access to table, column, and join key metadata through the TAP query mechanisms themselves. Users can discover tables or columns that meet their specific criteria by querying the tables described below. The service may enhance

Page : 22 sur 46 Zoom automatique

There are several approaches to getting metadata for a given TAP service. All TAP services **must** support a set of tables in a schema named **TAP_SCHEMA** that describe the tables and columns included in the service. In addition to the **TAP_SCHEMA**, there are two other ways to get metadata from a TAP service. First, the VOSI tables resource provides metadata on all tables and columns; this resource is described in 2.2.5 . The VOSI tables resource provides the same metadata as the **TAP_SCHEMA** but in a rigorously controlled format; the information in the **TAP_SCHEMA** is equivalent to that defined by the VODataService [7]. Second, the client may specify a query of one or more tables setting the **MAXREC** parameter to 0 so that only the metadata regarding the requested fields is returned. Use of **MAXREC** is described in 2.3.7 .

The **TAP_SCHEMA** provides access to table, column, and join key metadata through the TAP query mechanisms themselves. Users can discover tables or columns that meet their specific criteria by querying the tables described below. The service may enhance the **TAP_SCHEMA** with additional metadata where that seems appropriate; since it is self-describing, the **TAP_SCHEMA** may be queried to determine if any extended schema metadata is defined by the service. Services **must** provide these tables and make them accessible by all supported query mechanisms.

The qualified names in the tables of the TAP schema **must** follow the rules defined in section 2.4. The names **must** be stated in a form that is acceptable as an operand of a query.

- 21 -

Table Access Protocol

All columns in the **TAP_SCHEMA** tables are of type VARCHAR except for size, principal, indexed, and std (in Columns) which are INTEGER values.

List of fragments matching the query

Interop- Paris - Laurent Michel - May 2019

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

Step 4: Web Application

- Question: Is the TAP_SCHEMA mandatory?

Google-like question documents

TAP_SCHEMA must

Highlight Group By Documents Match Words Match Sentence Edit Query

5 matche(s)

REC-TAP-1.0
<http://saada.unistra.fr/esdoc/corpus//IVOA/REC-TAP-1.0.pdf>
Show Text - 11.238452

... **must** support a set of tables in a schema named **TAP_SCHEMA** that describe the tables and columns included in the service. In addition to the **TAP_SCHEMA** there are two other ways to get metadata from a TAP service. First, the VOSI tables resource provides metadata on all tables and columns; this resource is described in 2.2.5 . The VOSI tables resource provides the same metadata as the **TAP_SCHEMA** but in a rigorously controlled format; the information in the **TAP_SCHEMA** is equivalent to that defined by the VODataService...

REC-TAP-1.0
<http://saada.unistra.fr/esdoc/corpus//IVOA/REC-TAP-1.0.pdf>
Show Text - 10.575806

... **TAP_SCHEMA** provides access to table, column, and join key metadata through the TAP query mechanisms themselves. Users can discover tables or columns that meet their specific criteria by querying the tables described below. The service may enhance

There are several approaches to getting metadata for a given TAP service. All TAP services **must** support a set of tables in a schema named **TAP_SCHEMA** that describe the tables and columns included in the service. In addition to the **TAP_SCHEMA**, there are two other ways to get metadata from a TAP service. First, the VOSI tables resource provides metadata on all tables and columns; this resource is described in 2.2.5 . The VOSI tables resource provides the same metadata as the **TAP_SCHEMA** but in a rigorously controlled format; the information in the **TAP_SCHEMA** is equivalent to that defined by the VODataService...

Search keywords highlighted

Setting the **MAXREC** parameter to 0 so that only the metadata regarding the requested fields is returned. Use of **MAXREC** is described in 2.3.7 .

The **TAP_SCHEMA** provides access to table, column, and join key metadata through the TAP query mechanisms themselves. Users can discover tables or columns that meet their specific criteria by querying the tables described below. The service may enhance the **TAP_SCHEMA** with additional metadata where that seems appropriate; since it is self-describing, the **TAP_SCHEMA** may be queried to determine if any extended schema metadata is defined by the service. Services **must** provide these tables and make them accessible by all supported query mechanisms.

The qualified names in the tables of the TAP schema **must** follow the rules defined in section 2.4. The names **must** be stated in a form that is acceptable as an operand of a query.

Table Access Protocol

All columns in the **TAP_SCHEMA** tables are of type VARCHAR except for size, principal, indexed, and std (in Columns) which are INTEGER values.

Original text positioned on the select fragment

Thank You

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