
CERN

# Modernising CERN Document Conversion service

Ruben Gaspar, IT-CDA-IC

# Agenda

- What this service is about
- Old service
- New service

# Aim/Mandate

The mandate of the document Conversion Service is to provide automatic conversion of documents, usually office documents, towards a different format usually PDF or PDF/A. This is accomplished nowadays using a REST API. Workflow is asynchronous.

Evolution to different formats (input or output) is possible.

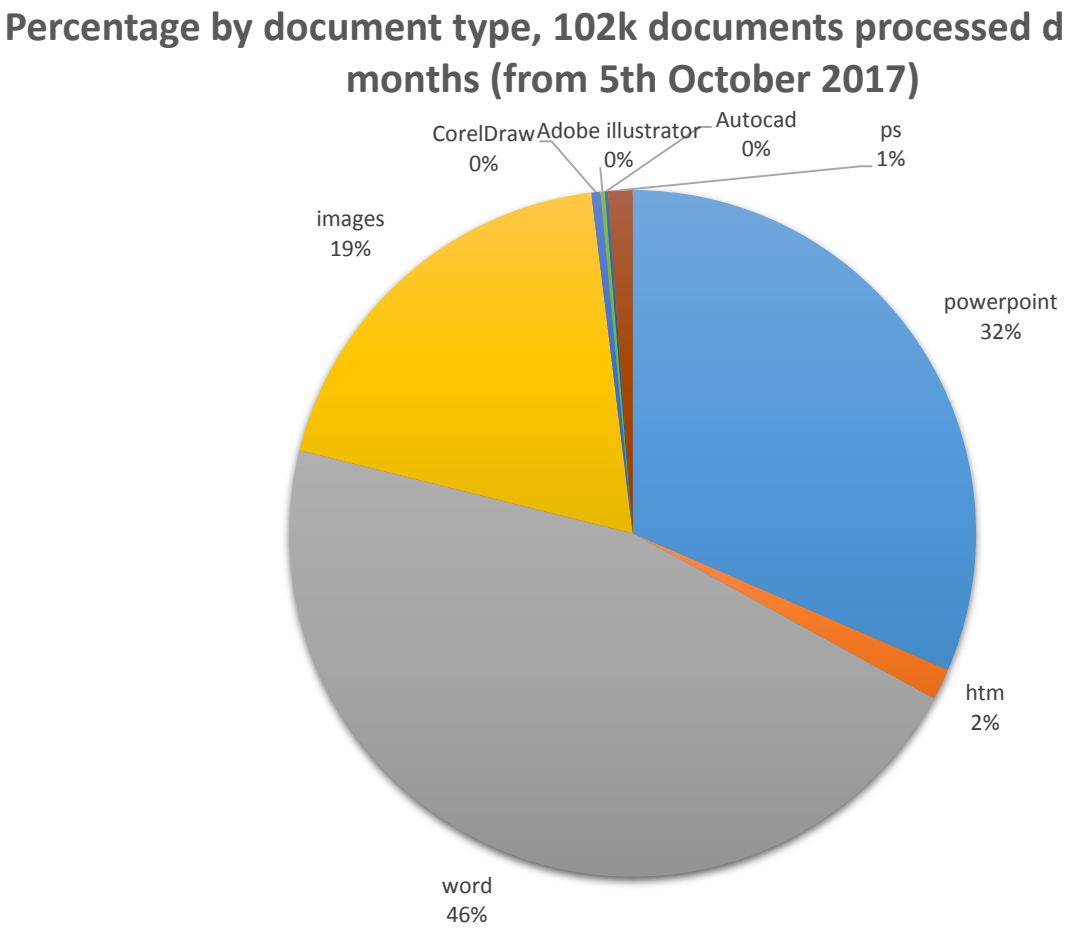
# Agenda

- What this service is about
- Old service
- New service

# Old converter

- In operation since 10 years
- Multithreading application but only works with 1 thread
- Programmed in Python2
- Missing any coding convention
- Missing any structure
- Very reduced logging and no testing framework
- Based on DFS as a backend to coordinate jobs and store input and output documents
- Running on Microsoft IIS server on the worker nodes (no load balancing)

# Statistics




Indico & EDMS workload

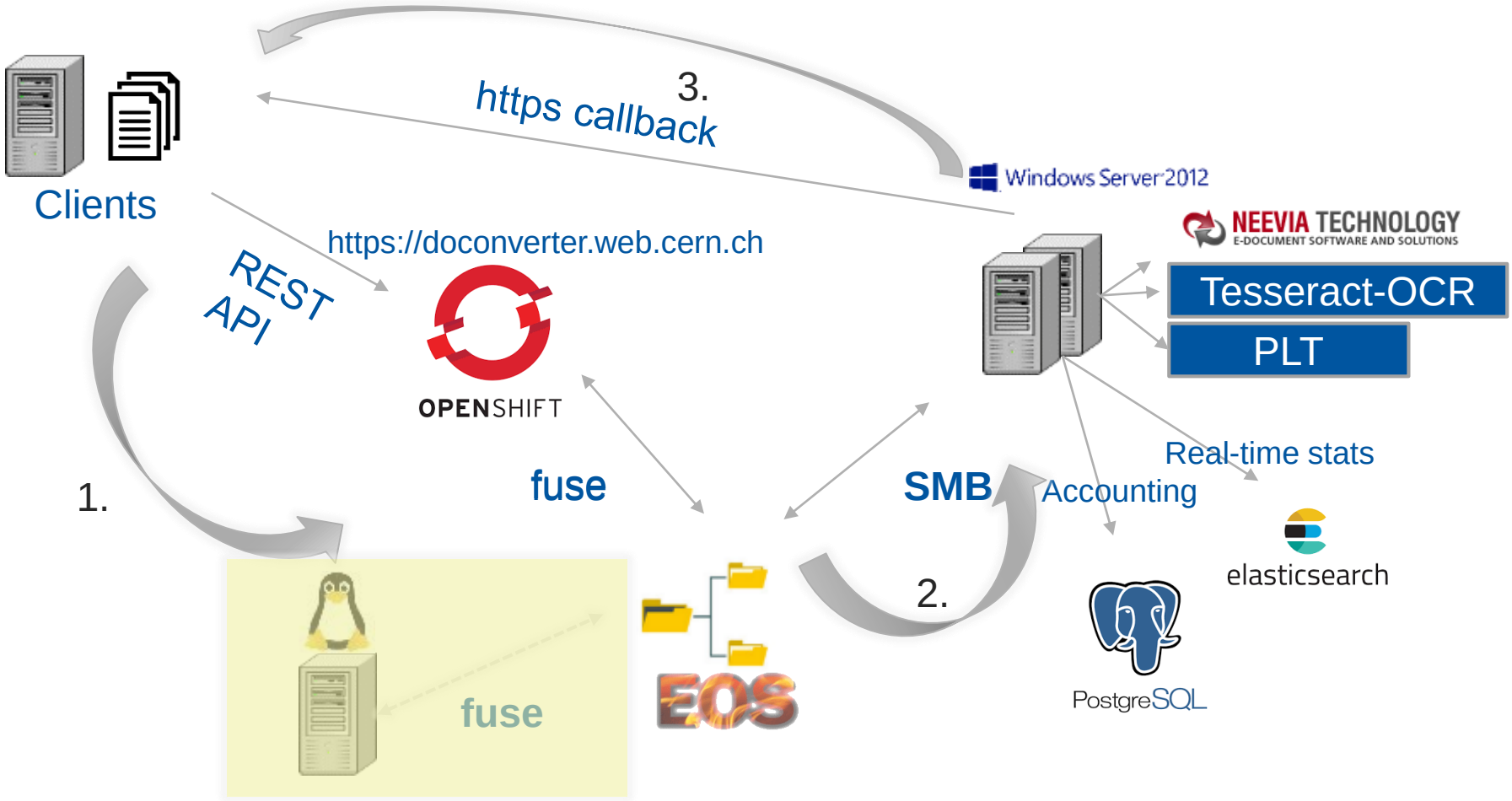
# Agenda

- What this service is about
- Old service
- New service

# New service guidelines design

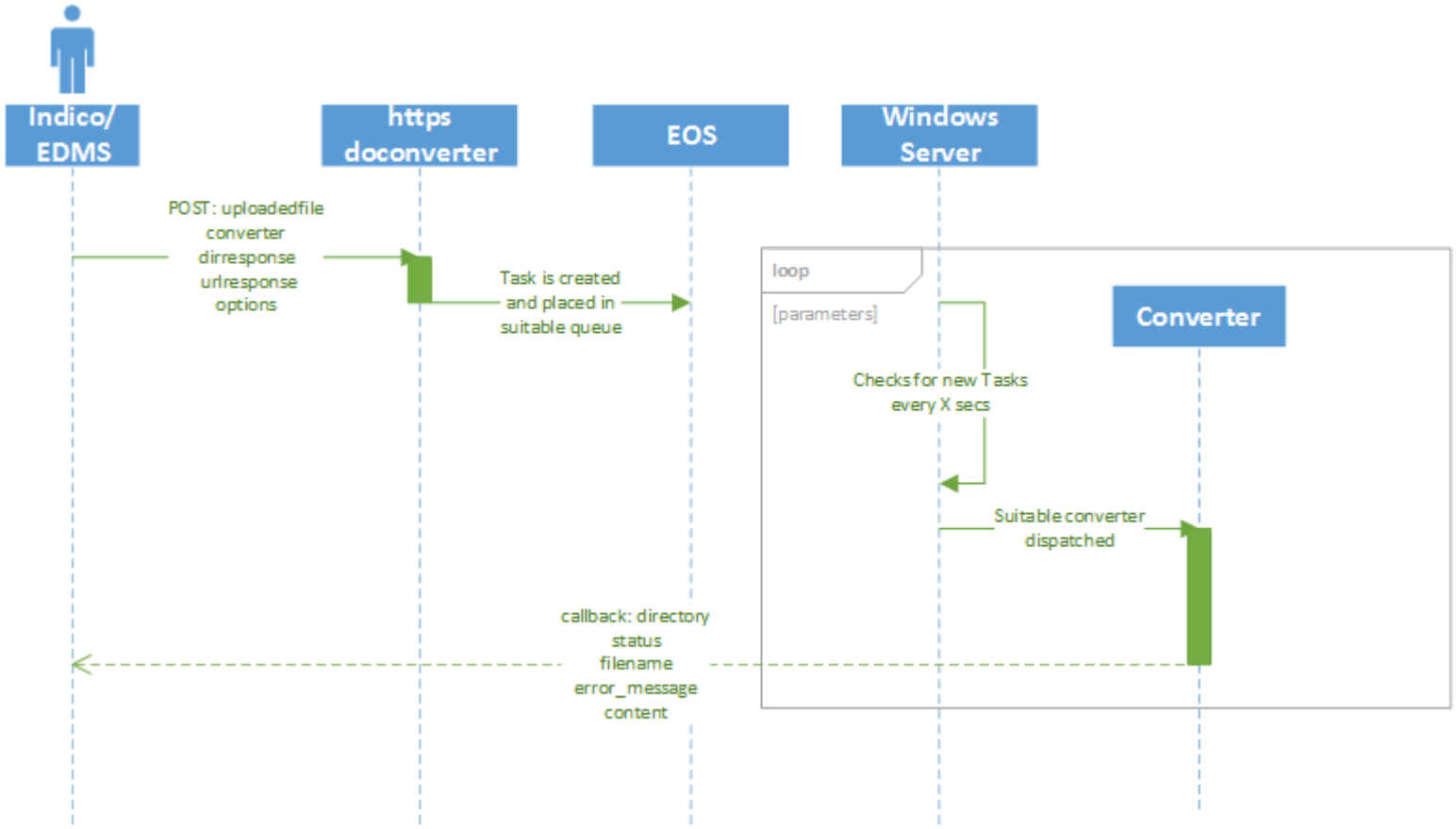
- Keep as much as possible same interface with user community and converters admins
- Re-design program architecture
  - Object Oriented
- Adding new conversions should be as simple as possible
  - Exe type: HPGL converter, Tesseract-ocr
  - COM type: Neevia
  - REST type
- Renew technology: using IT services as much as possible
  - IT services: ES, Openstack, Openshift, EOS & DBoD

# New architecture




Possible in the future!

# Accessing the service



# New Doconverter

- OSS: code at Github[1]
- Developed in Python3
  - Flask, Flask-sqlalchemy,…
- Multi-process application
  - 50% faster than old system while processing same workload
- Scalability/redundancy by adding more worker nodes or using some features of the infrastructure e.g. Openshift autoscale, HAProxy
- Doc based on gitbook[2]

[1] https://github.com/CERNCDAIC/doconverter
[2] https://cern.ch/docdocs or https://cernbox.cern.ch/index.php/s/Sb8lLIVgMBwqSzR (PDF book)

# doconverter.ini: how to customize

```
[default]
extensions_all=doc,docx,ppt,pptx,xlsx,tif,htm,txt,png,jpg,pdf,plt
prefix_dir=Y:\
archival_dir=Y:\
ca_bundle=c:\doconverter\cert\COMODO_OV_SHA-256_bundle.crt
servers=doconvert01,doconvert02
doconvert01=doc,docx,ppt,pptx,xlsx,tif,htm,txt,png,jpg,pdf,ps,pdfa,thumb,toimg,plt,hpgl,tesocr,modiocr
doconvert02=doc,docx,ppt,pptx,xlsx,tif,htm,txt,png,jpg,pdf,ps,pdfa,thumb,toimg,plt,hpgl,tesocr
[manager]
converters=Neevia,Hpglview_raster,Tesseract_ocr
stopper=c:\doconverter\noconverter.txt
[monitor]
emails=admin@cern.ch
tasksalert=50
smtpserver=XXXXX.cern.ch
[Neevia]
extensions_allowed=doc,docx,ppt,pptx,xlsx,tif,htm,txt,png,jpg,pdf
output_allowed=pdf,png,ps,pdfa,thumb,toimg,modiocr
type=windows
exe=dConverter.exe
[Hpglview_raster]
….
[database]
host=XXXXX.cern.ch
port=6607
db=dbconverter
user=postgresql://dbconverter
password=xxxxx
[test]
url=https://xxxxx/doconverter/api/v1.0/uploads
url_response=http://XXXX/doconverter/api/v1.0/received
diresponse=c:\doconverter\testing
files=c:\doconverter\files\wordfile.docx,c:\doconverter\files\excelfile.xlsx,c:\doconverter\files\htmfile.htm
```

Accepted input extensions

Accepted input and output extensions by a server. Asymmetric configuration possible

Converters configuration

logging

Functional testing

# Possible conversions

NEW!

| Input File | Conversion output | Options | Expected result | Comments |
|---|---|---|---|---|
| Office 2016, Autocad(cdx), OpenOffice, Coreldraw(cdr),rtf,htm,bmp,jpg,tif | converter=[pdf,dfa,ps] | N/A | pdf file | N/A |
| Special case for Word documents(e.g. doc,docx) | converter=[pdf,pdfa] | hidedocumentrevisions=[false,true] | pdf file | You may want to have comments visible on your pdf, by default they are not |
| PDF file | converter=toimg | typeofimg=[jpeg,bmp,tiff,png]:imgresh=200:imgresv=200* | zip or tif file | Depending on the format of file chosen a tif file or zip file containing all the numbered pages of the document will be sent back. E.g. original file mypresentation.pdf -> zip: mypresentation.zip or tif: mypresentation.tif |
| PDF file | converter=thumb | imgresh=200:imgresv=200:imgheight=300:imgwidth=300** | png file | if original file was called mydocument.pdf -> mydocument1.png |
| PLT file | converter=hpgl | color=[true,false] | pdf file | N/A |
| tif,png,jgp | converter=tesocr | N/A | pdf searchable file | It uses tesseract-ocr as engine that just support image files |
| pdf, tif | converter=modiocr | language=[english,french] | pdf searchable file | N/A |

*__imgresh,imgresv__: should be one of
*'72x72','100x100','150x150','200x200','300x300','400x400','600x600','1200x1200'*
**__imgresh,imgresv,imgheight,imgwidth__:should be one of
*'72x72','100x100','150x150','200x200','300x300','400x400','600x600','1200x1200'* and *imgheight_imgwidth*
should be an integer in pixels e.g.: thumb_200_200_150_150

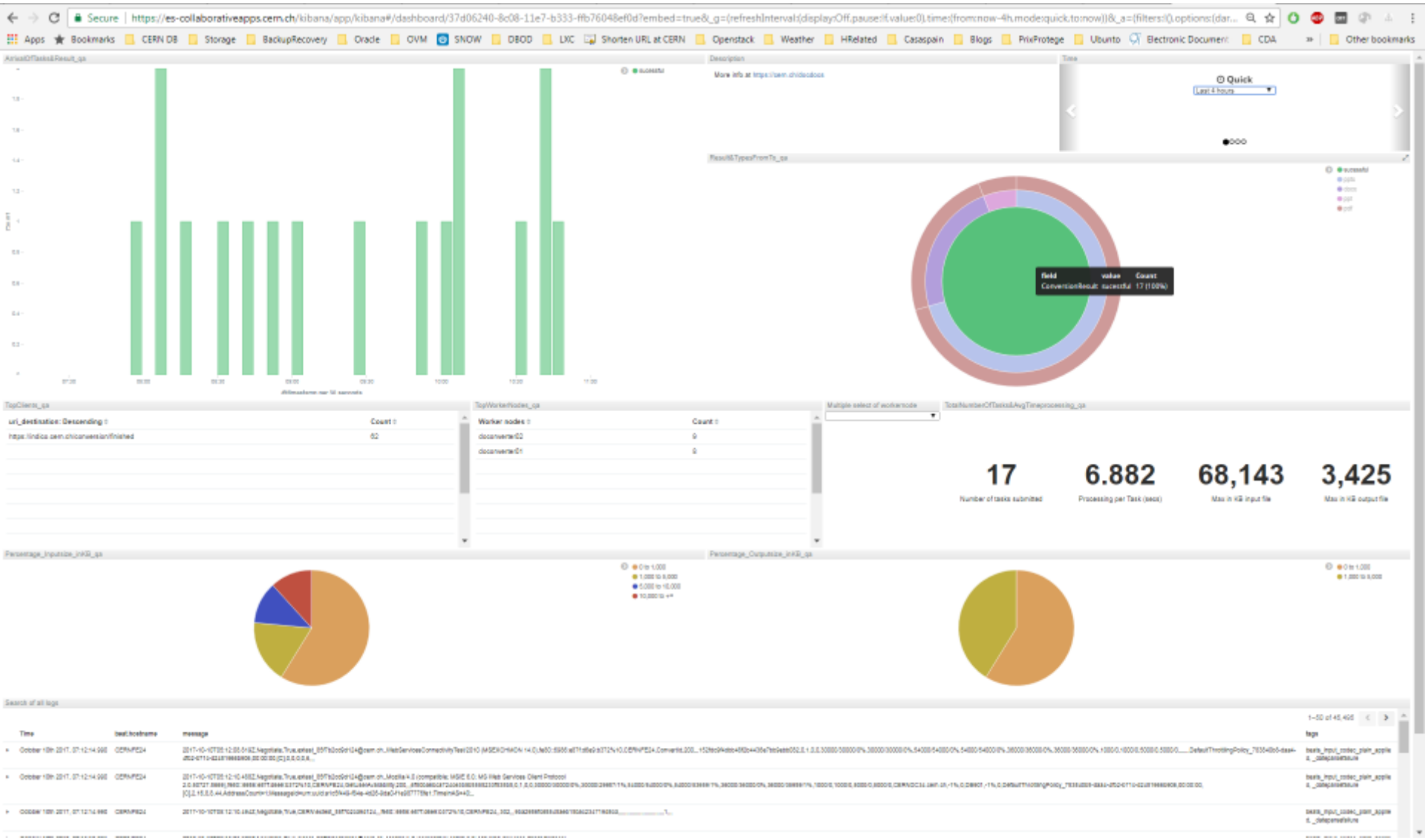
From documentation:
https://cernbox.cern.ch/index.php/s/Sb8LIVgMBwqSzR

# Real time stats: elasticsearch

- ES cluster to analyse and extract some info from the document converter servers logs
- Puppet module[1] to configure logstash
  - Done to scale: being used by other services: E-mail, AVC
  - SSL enforced through out the whole network path

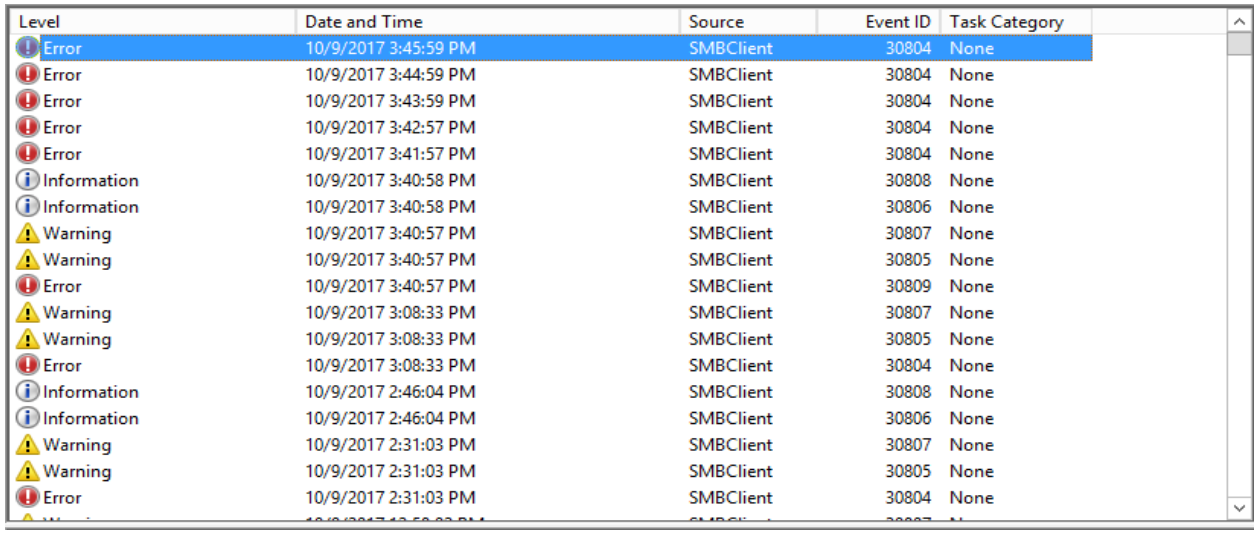
[1] https://gitlab.cern.ch/ai/it-puppet-hostgroup-doconverter

# Dashboard view



# EOS as a file system backend

- Allows to keep state among the different players
- Some instabilities while accessing it via SMB
  - Following that with EOS and Network support



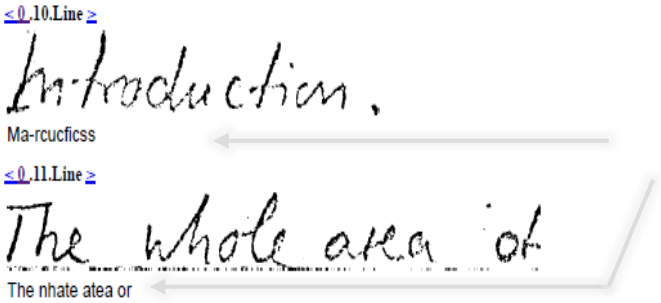
| Level | Date and Time | Source | Event ID | Task Category |
|---|---|---|---|---|
| Error | 10/9/2017 3:45:59 PM | SMBClient | 30804 | None |
| Error | 10/9/2017 3:44:59 PM | SMBClient | 30804 | None |
| Error | 10/9/2017 3:43:59 PM | SMBClient | 30804 | None |
| Error | 10/9/2017 3:42:57 PM | SMBClient | 30804 | None |
| Error | 10/9/2017 3:41:57 PM | SMBClient | 30804 | None |
| Information | 10/9/2017 3:40:58 PM | SMBClient | 30808 | None |
| Information | 10/9/2017 3:40:58 PM | SMBClient | 30806 | None |
| Warning | 10/9/2017 3:40:57 PM | SMBClient | 30807 | None |
| Warning | 10/9/2017 3:40:57 PM | SMBClient | 30805 | None |
| Error | 10/9/2017 3:40:57 PM | SMBClient | 30809 | None |
| Warning | 10/9/2017 3:08:33 PM | SMBClient | 30807 | None |
| Warning | 10/9/2017 3:08:33 PM | SMBClient | 30805 | None |
| Error | 10/9/2017 3:08:33 PM | SMBClient | 30804 | None |
| Information | 10/9/2017 2:46:04 PM | SMBClient | 30808 | None |
| Information | 10/9/2017 2:46:04 PM | SMBClient | 30806 | None |
| Warning | 10/9/2017 2:31:03 PM | SMBClient | 30807 | None |
| Warning | 10/9/2017 2:31:03 PM | SMBClient | 30805 | None |
| Error | 10/9/2017 2:31:03 PM | SMBClient | 30804 | None |

- Using the sync client[1] from Windows worker nodes for a more stable solution → increase latency

[1] https://cernbox.cern.ch/cernbox/doc/clients.html

# OCR & ICR

- Number of programs tested (see next)
- A number of OCR solutions offered  good results
- ICR is still not mature




ICR view

- Resolution is key, at least 200x200 dpi
- Conversion Service offers two possible solution “modiocr” and “tesocr”

# OCR & ICR

| Product | OCR | ICR (handwriting) | remarks |
|---|---|---|---|
| PDF Editor 6 Professional | N/A | N/A | It didn't work either using Windows 7 or 2012R2. Support didn't provide any info |
| CVISION: PDFCompressor Professional Edition v6.5.1177 | Ok | Bad | Accepts PDF and images. Results as PDF or Word doc. Watch folder mechanism |
| http://www.i2ocr.com/ | Ok | Bad | Admits images, Not easy to work with. No RESTFul api |
| http://www.smartocr.com/ | Ok | Bad | Admits PDF. No RESTFul API but a exe program |
| A2iA: TextReader | OK | ~OK | It should work with proper handwritten docs and resolution |
| **Tesseract-OCR** | **OK** | **Bad** | **Open source, works with images only.** |
| **Microsoft MODI (Module of sharepoint designer)** | **OK** | **Bad** | **High integration with Neevia. It works with Image and PDF files.** |

# Conclusions

- Faster and more feature-rich system that relies on IT services
- Simpler to evolve the service, adding more conversions
  - E.g. posters (PDF/X)
- Possible to integrate with other services e.g. CERNBox

# Acknowledgements


- IT colleagues in general!
- And specially:
  - Thomas Baron (IT-CDA) for all Document Conversion support
  - Pablo Saiz (IT-CM) for his help on setting up ES cluster(s)
  - Alberto Rodriguez (IT-CDA) for all Openshift support
  - EOS support guys!