TEXT-BASED INFORMATION EXTRACTION SERVICES

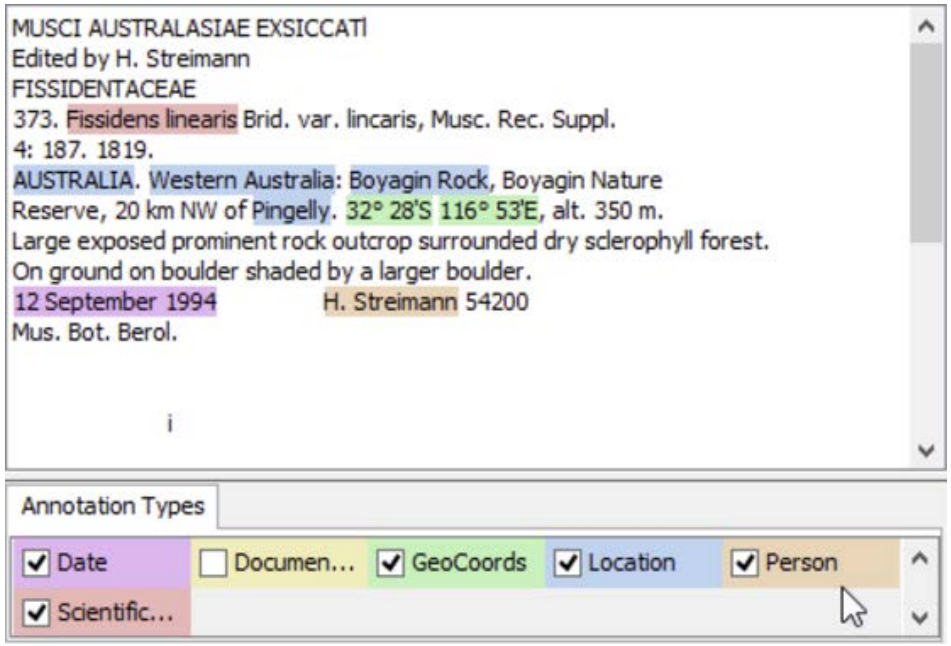
Dr. Eduard Santamaria eduard.santamaria@iosb.fraunhofer.de

Berlin, 1-2 June 2017

Fraunhofer IOSB

TEXT-BASED INFORMATION EXTRACTION SERVICES

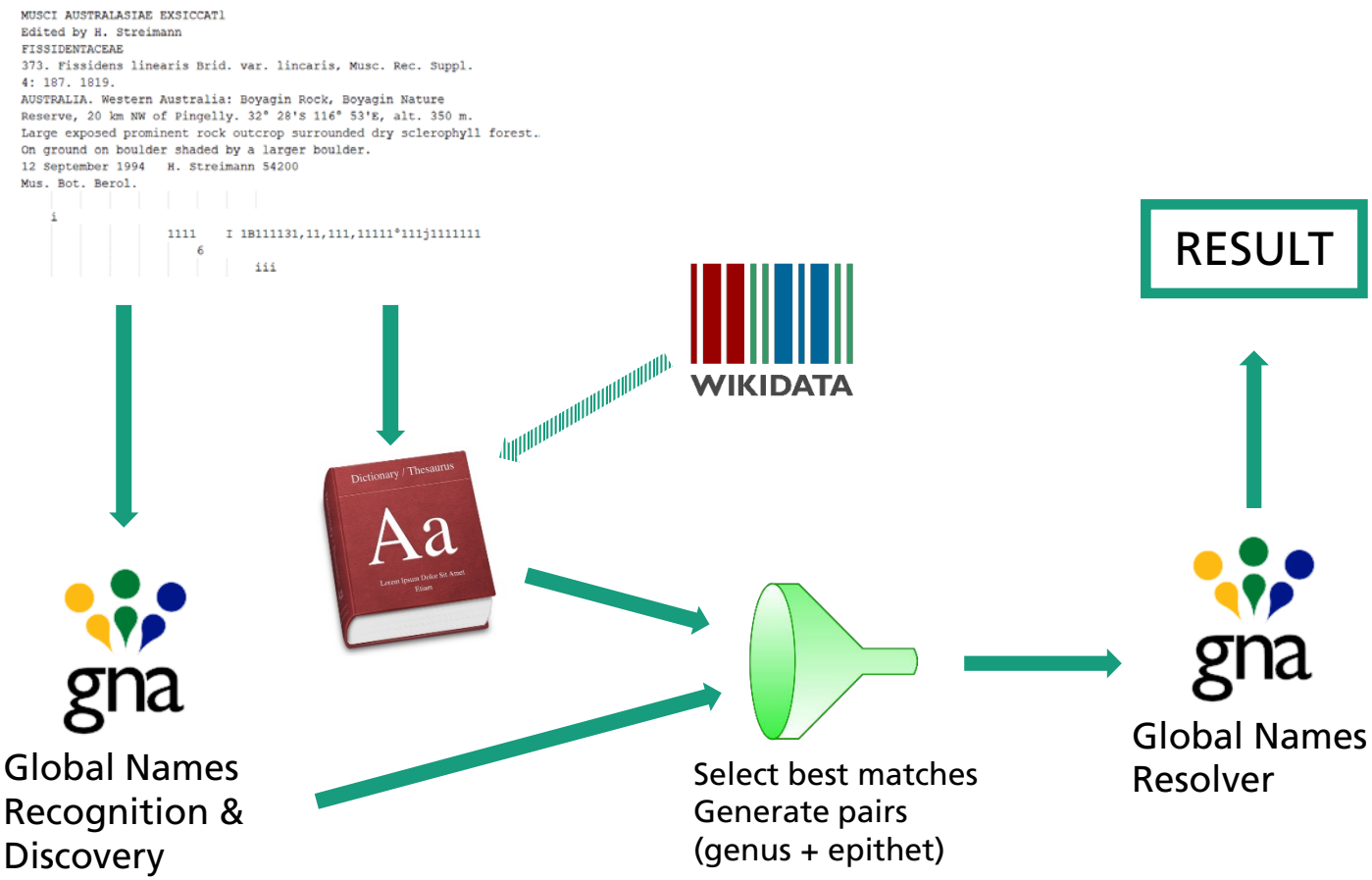
- Given a piece of text (not an image) with the contents of the specimen label, extract important metadata about the specimen:
  - Scientific name
  - Collector
  - Collection date
  - Geo-coordinates
  - Location



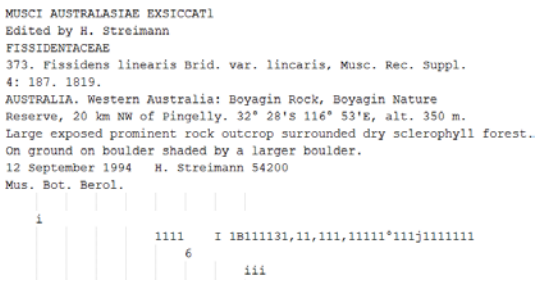
MAIN PROBLEMS

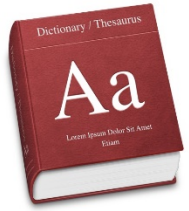
- Errors in text (OCR does not always work)
  - Deterioration of specimen sheet
  - Multiple languages (with particular characters)
  - Handwritten parts
- Syntax and abbreviations
  - Authors try to fit much information in a small label
- Ambiguity
  - Stone – family name or description of the place?
  - White – family name, color?

SCIENTIFIC NAME



COLLECTOR


MUSCI AUSTRALASIAE EXSICCATI
Edited by H. Streimann
FISSIDENTACEAE
373. Fissidens linearis Brid. var. lincaris, Musc. Rec. Suppl.
4: 187. 1819.
AUSTRALIA. Western Australia: Boyagin Rock, Boyagin Nature
Reserve, 20 km NW of Pingelly. 32° 28'S 116° 53'E, alt. 350 m.
Large exposed prominent rock outcrop surrounded dry sclerophyll forest.
On ground on boulder shaded by a larger boulder.
12 September 1994 H. Streimann 54200
Mus. Bot. Berol.



Harvard University Herbaria

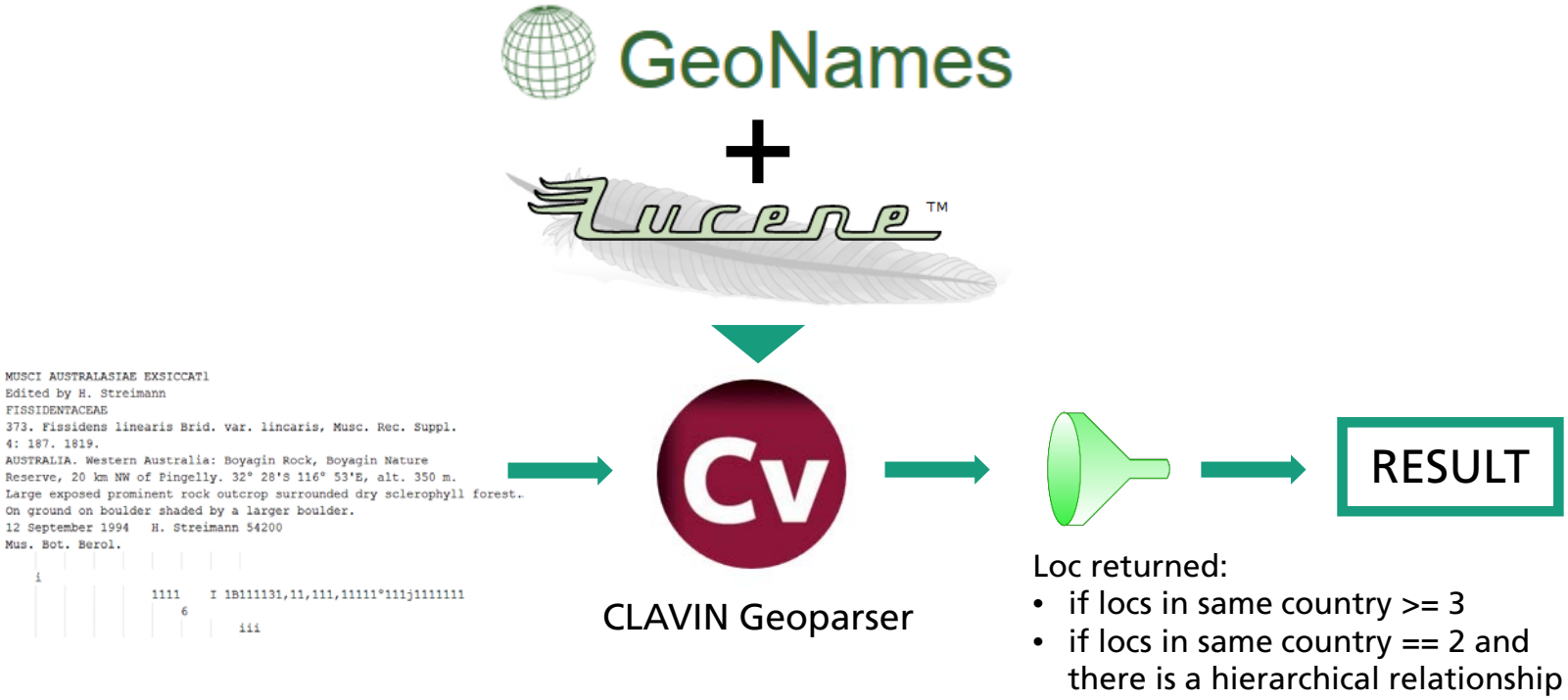
Best matches

The International Plant Names Index

Best matches
Take into account multiple forms:
Jan-Peter Frahm
J.-P. Frahm
Frahm, J.-P.
Accompanied by Leg., Det. or equivalent form?

RESULT

LOCATION

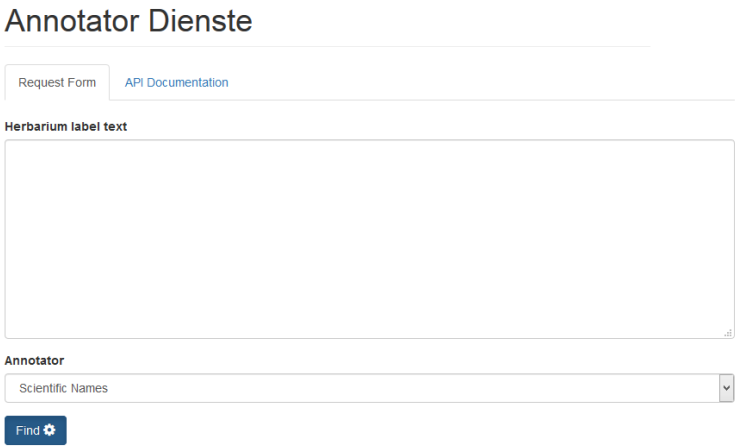

GeoNames
+
Lucene™
MUSCI AUSTRALASIAE EXSICCATI
Edited by H. Streimann
FISSIDENTACEAE
373. Fissidens linearis Brid. var. lincaris, Musc. Rec. Suppl.
4: 187. 1819.
AUSTRALIA. Western Australia: Boyagin Rock, Boyagin Nature
Reserve, 20 km NW of Pingelly. 32° 28'S 116° 53'E, alt. 350 m.
Large exposed prominent rock outcrop surrounded dry sclerophyll forest.
On ground on boulder shaded by a larger boulder.
12 September 1994 H. Streimann 54200
Mus. Bot. Berol.
i
1111 I 1B111131,11,111,11111°111j1111111
6
iii
Cv
CLAVIN Geoparser
RESULT
Loc returned:
• if locs in same country >= 3
• if locs in same country == 2 and there is a hierarchical relationship

DATE AND GEO-COORDINATES

- Date and Geo-coordinates services are based on regular expressions.
  - "(\\d{1,2})\\.\\s(\\w+)\\s(\\d{4})" → 5. VII 1875
  - "(\\w+)[\\,|\\.]?\\s(\\d{1,2})[\\.|\\,]\\s+(\\d{4})" → Mayo, 21. 2000
  - "(\\-?\\d{1,3})\\s?[°d*]?\\s?(\\d{1,3})(?:\\'|\\´)\\s?(\\d{1,3})(?:\\"|\\'')\\s?([nsewNSEW])“ → 47° 16'39" E
  - etc.
- Multiple languages for months:
  - English, English short form, French, Spanish, German, Italian, Portuguese
- Numbers: 6, 06, vi, but also v1 and o6

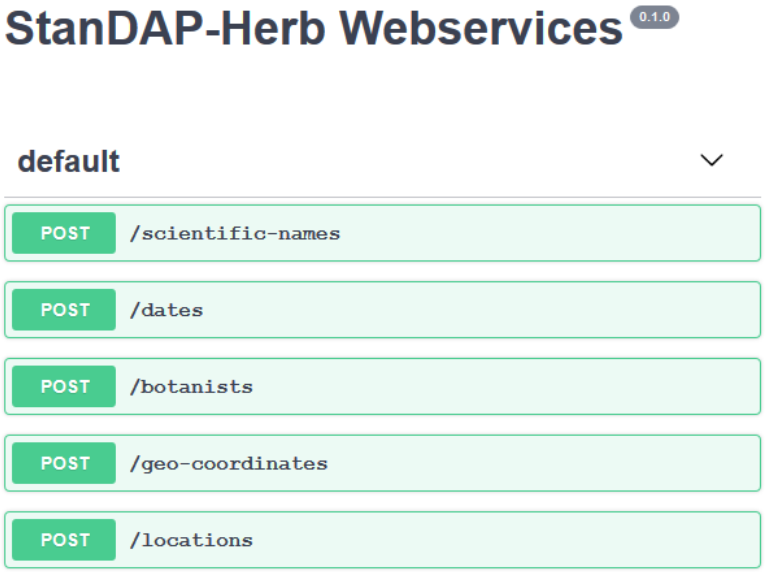
USAGE OPTIONS

- Web User Interface



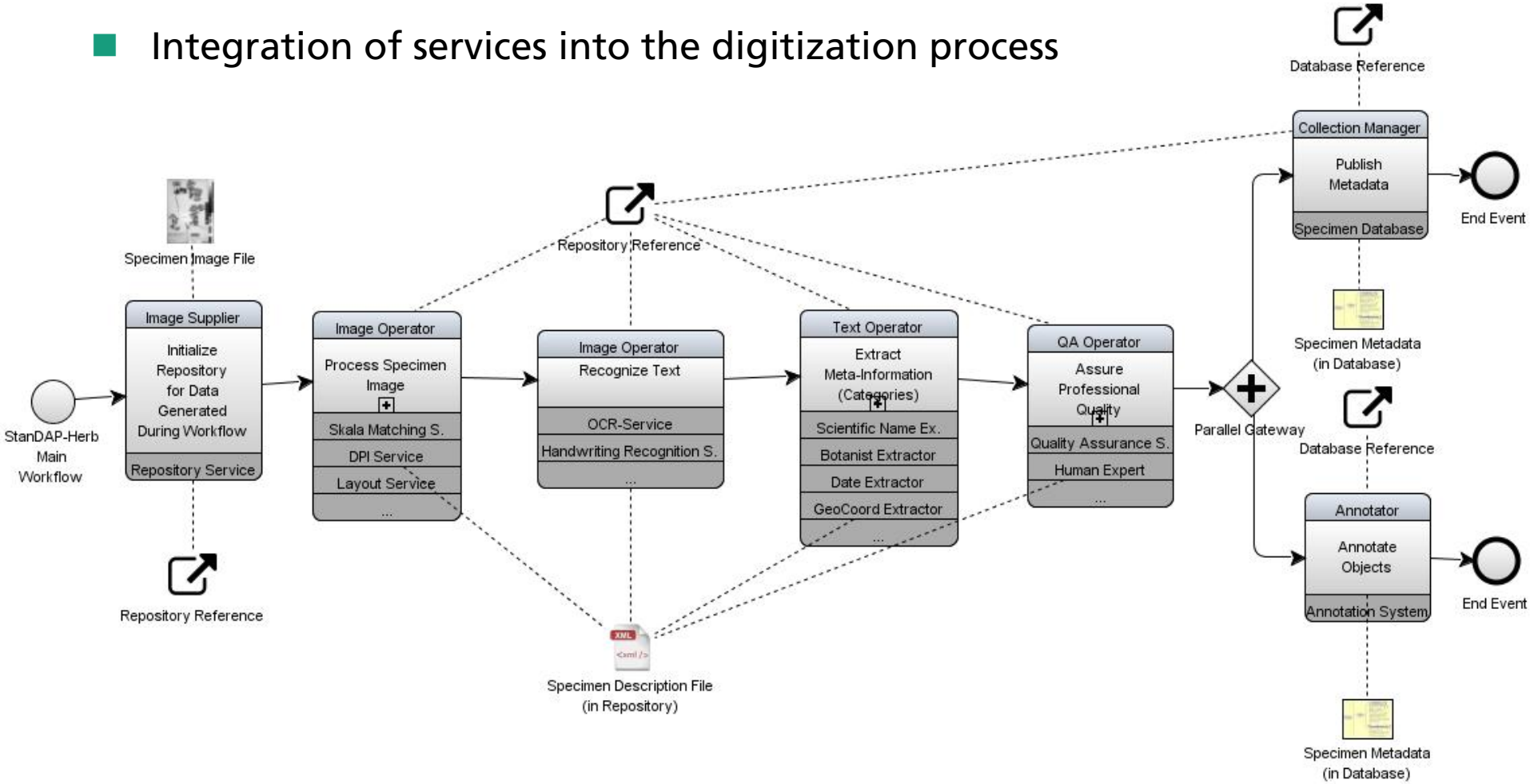
- Java Library

- REST Web Service



WORKFLOW

- Integration of services into the digitization process



RESULTS WITH TEST DATASET (I)

1162 text files with output from OCR.

Precision & Recall values automatically computed comparing service results with known correct results.

	Precision	Recall
Scientific Name	0,50	0,48
Collector Name	0,73	0,23
Date	0,66	0,50
GeoCoordinates	0,48	0,33
Country Name*	0,85	0,54

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

* Service provides other locations but only country was compared.

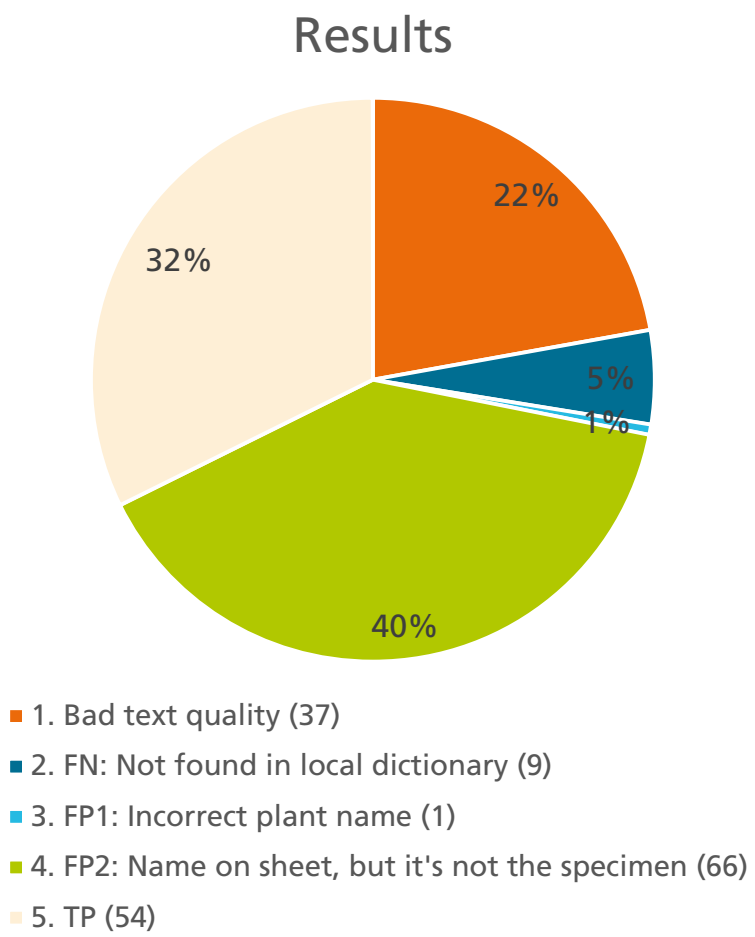
Fraunhofer IOSB

RESULTS WITH TEST DATASET (II)

- 100 random files
- 100 plant names

If OCRed text is ok, scientific name is found most of the time.

A few sheets contain lots of plant names



Fraunhofer IOSB

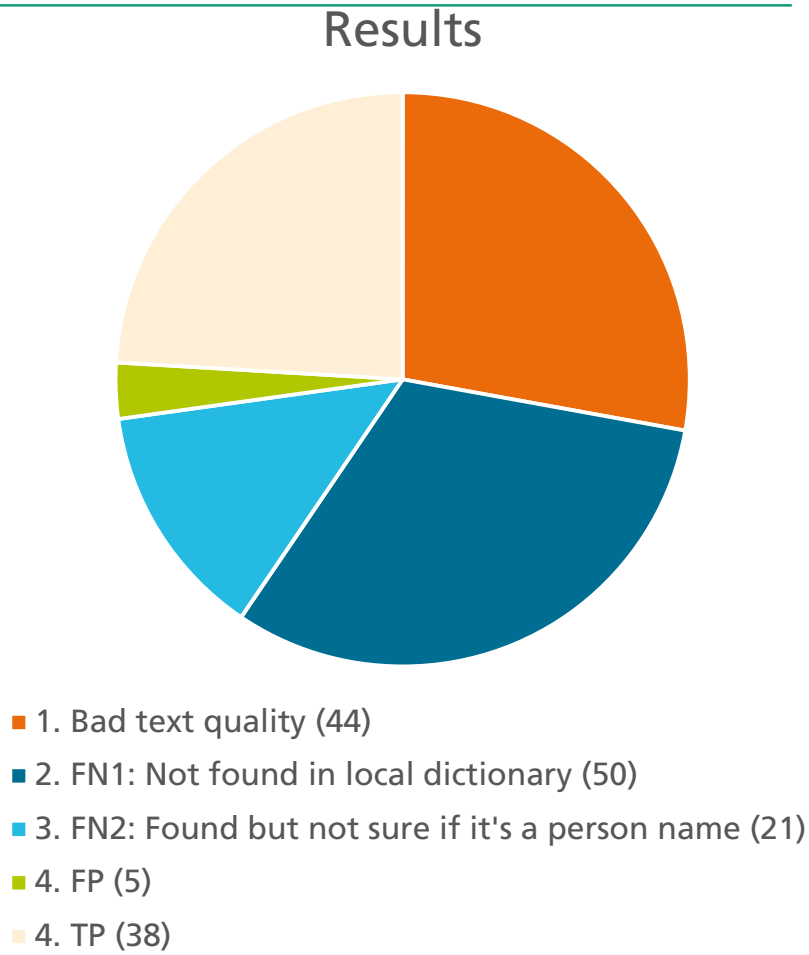
RESULTS WITH TEST DATASET (III)

- 100 random files
- 153 collector names

Number of good results limited by:

- Text not good enough for the service.
- Extent of dictionaries *
- Limitations of the annotator or ambiguity (ex. Stone)

* Remote dictionary queried only if local hit found.



Fraunhofer IOSB

POTENTIAL IMPROVEMENTS

- Automatically update local dictionaries periodically.
- Extend local dictionaries with input from users as specimen sheets are processed.
- Use statistical indicators to select candidates:
  - Which one of a number of candidates did the user more often select as the correct answer?
- Currently, the information extraction services work in isolation:
  - Date plus collector name could be checked with an external database to determine location.

TEXT-BASED INFORMATION EXTRACTION SERVICES