Semantic PDF Segmentation for Legacy Documents in Technical Documentation

Jan Oevermann

jan.oevermann@dfki.de

German
Research Center
for Artificial
Intelligence

SEMANTiCS 2018, Vienna, 13.09.18

Technical Documentation

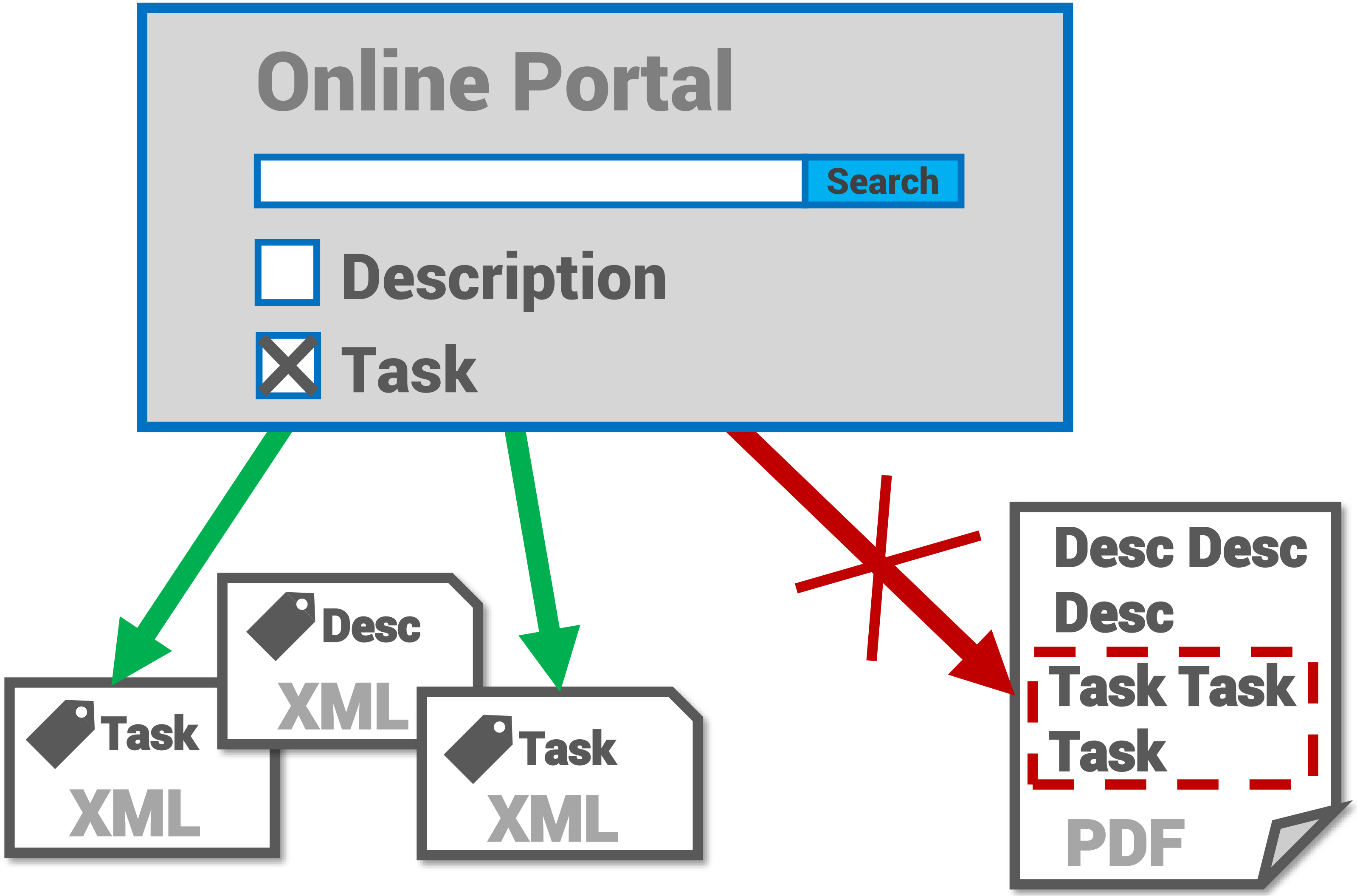
Most common: **PDF** documents

- “Digital Paper”, archival & distribution
- ISO Standard, guaranteed reproduction, ubiquitous support

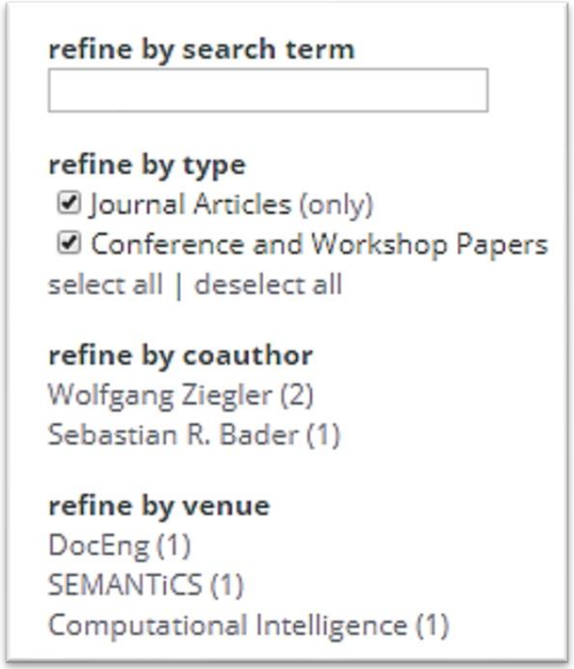
Best practice: **XML** content components

- Self-contained building blocks, e.g. chapter-sized, ~150-500 words
- Reuse, translation, aggregation, delivery

Motivation


Online Portal
Search
□ Description
☒ Task
Desc
XML
Task
XML
Task
XML
Desc Desc
Desc
Task Task
Task
PDF

Motivation



refine by search term

refine by type

☑ Journal Articles (only)

☑ Conference and Workshop Papers

select all | deselect all

refine by coauthor

Wolfgang Ziegler (2)

Sebastian R. Bader (1)

refine by venue

DocEng (1)

SEMANTiCS (1)

Computational Intelligence (1)

Only **safety information** of the document

I need **maintenance information** about the **fuel injection**

Everything about the **hydraulic pump** in **technical overview** or **technical data**

Faceted search

Information request with semantic concepts which can be used as facets

Motivation

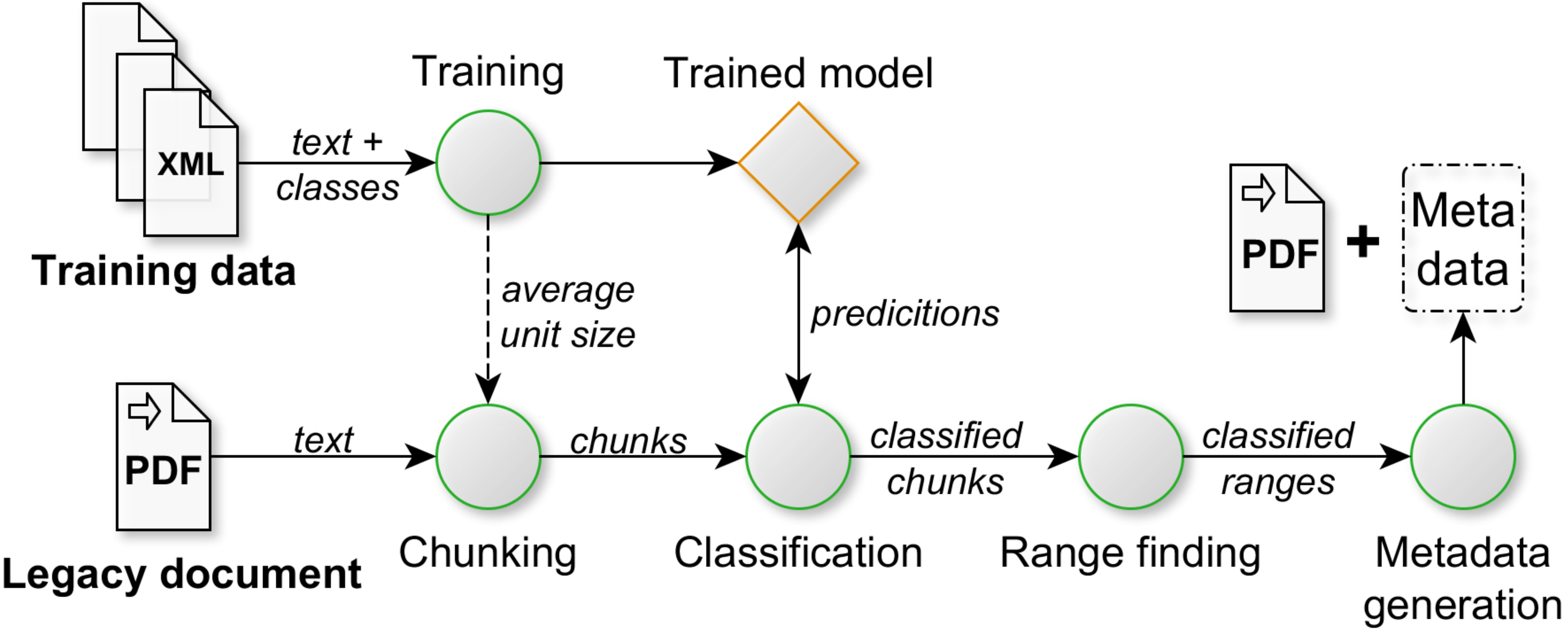
Limitations of PDF

- Semantic structure gets lost
- No metadata for (overlapping) segments
- Large documents (>200p) only accessible via full text search

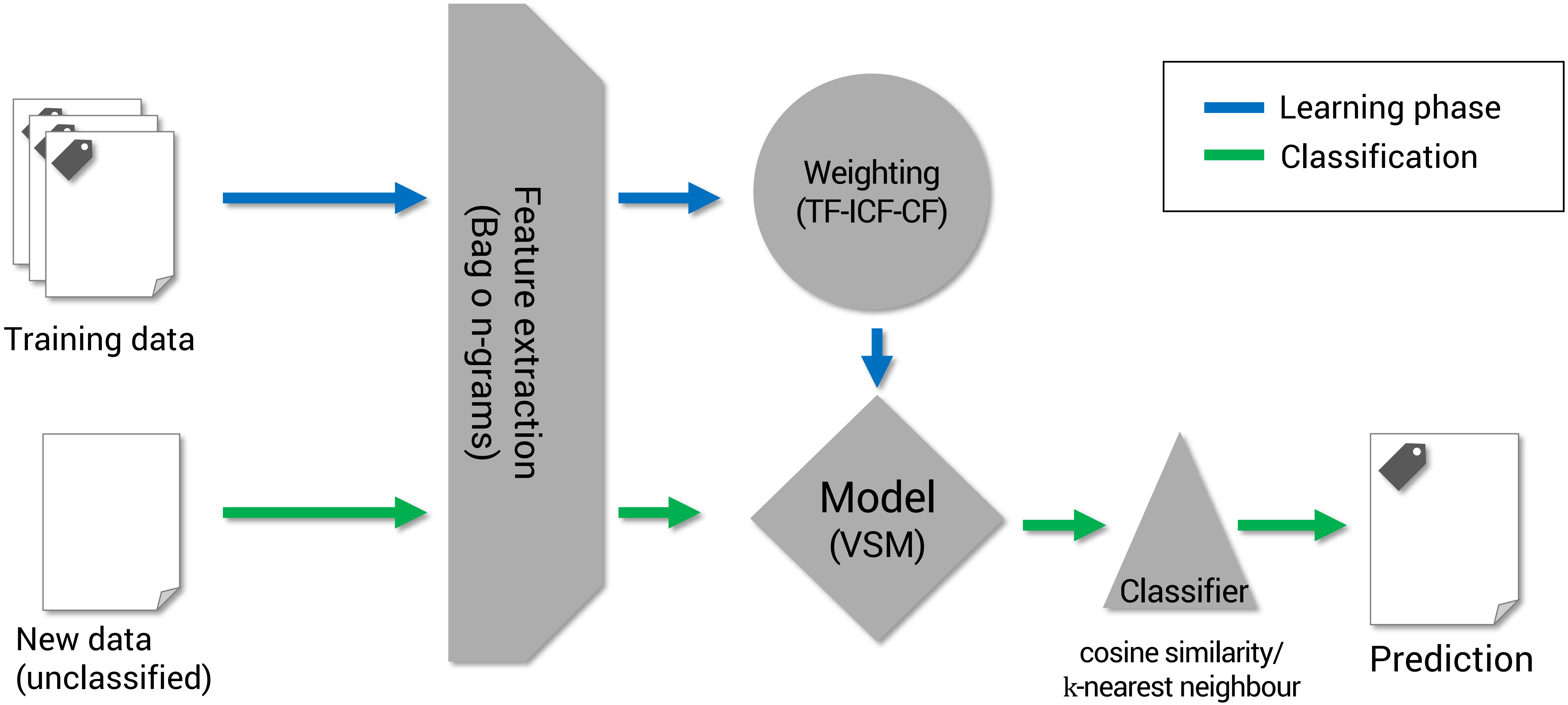
Idea

- Use knowledge from structured XML content components
  - Manually annotated semantic concepts / metadata
- Apply trained model on text extracted from PDF
- Find segments which are semantically relevant

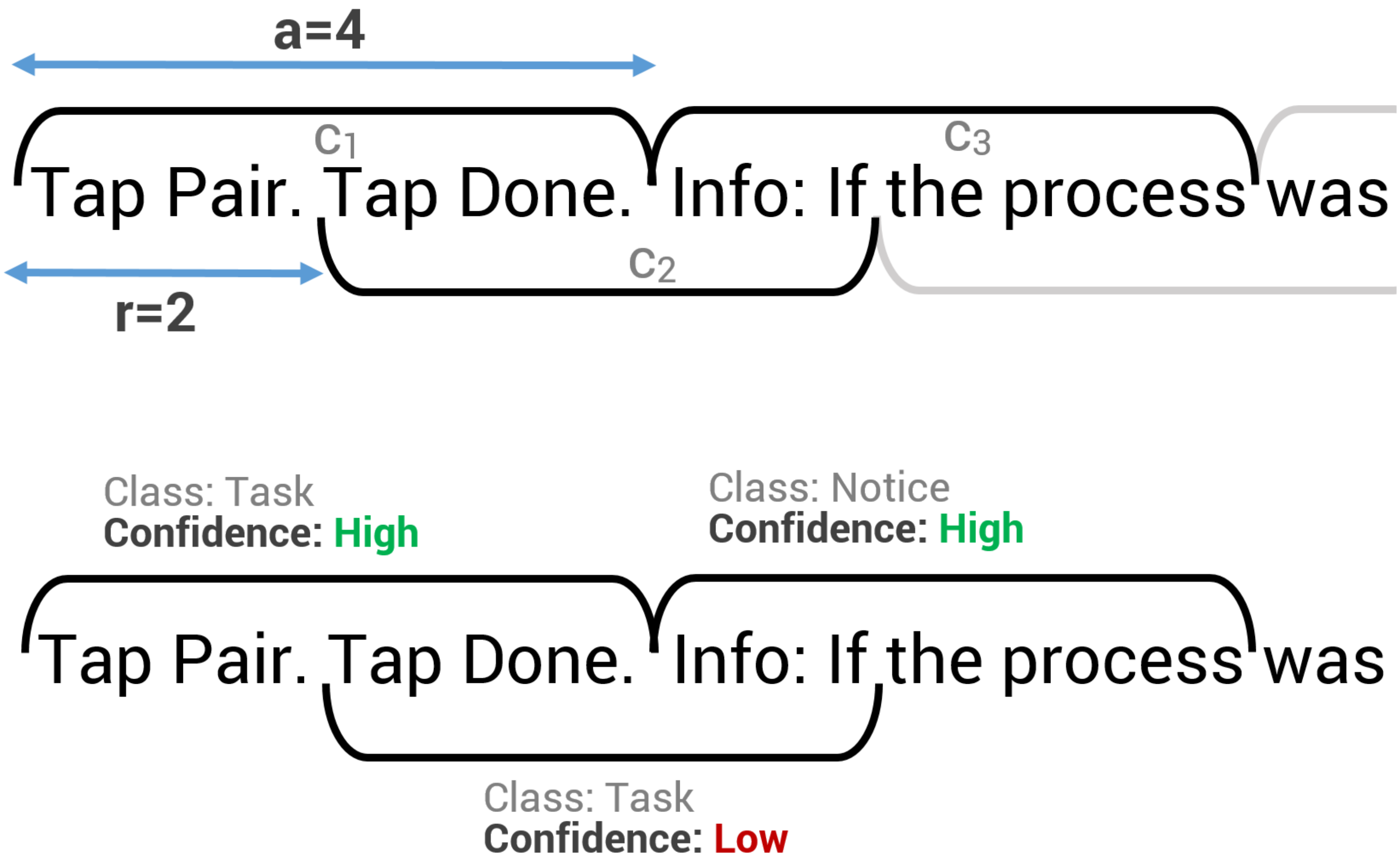
Procedure model



Training / Classification



Chunking



a=4

c_1 c_3

Tap Pair. Tap Done. Info: If the process was

c_2

r=2

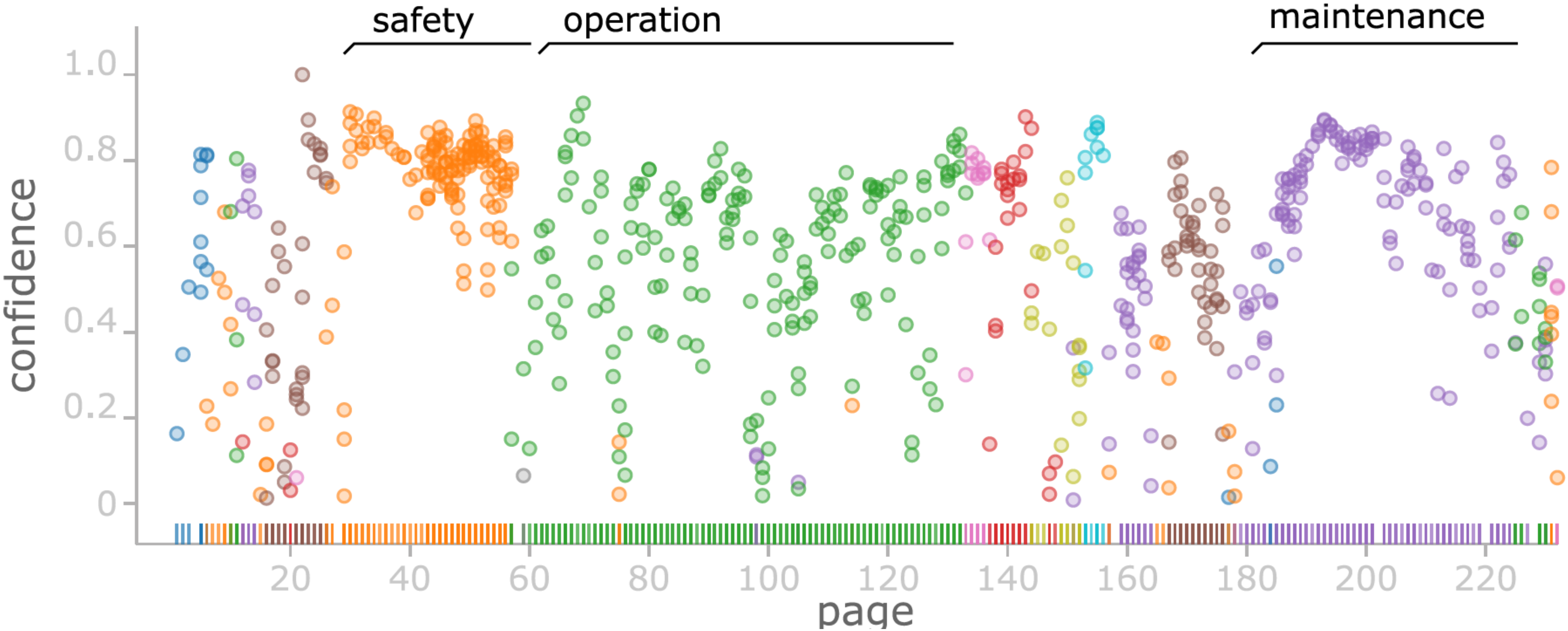
Class: Task
Confidence: High

Class: Notice
Confidence: High

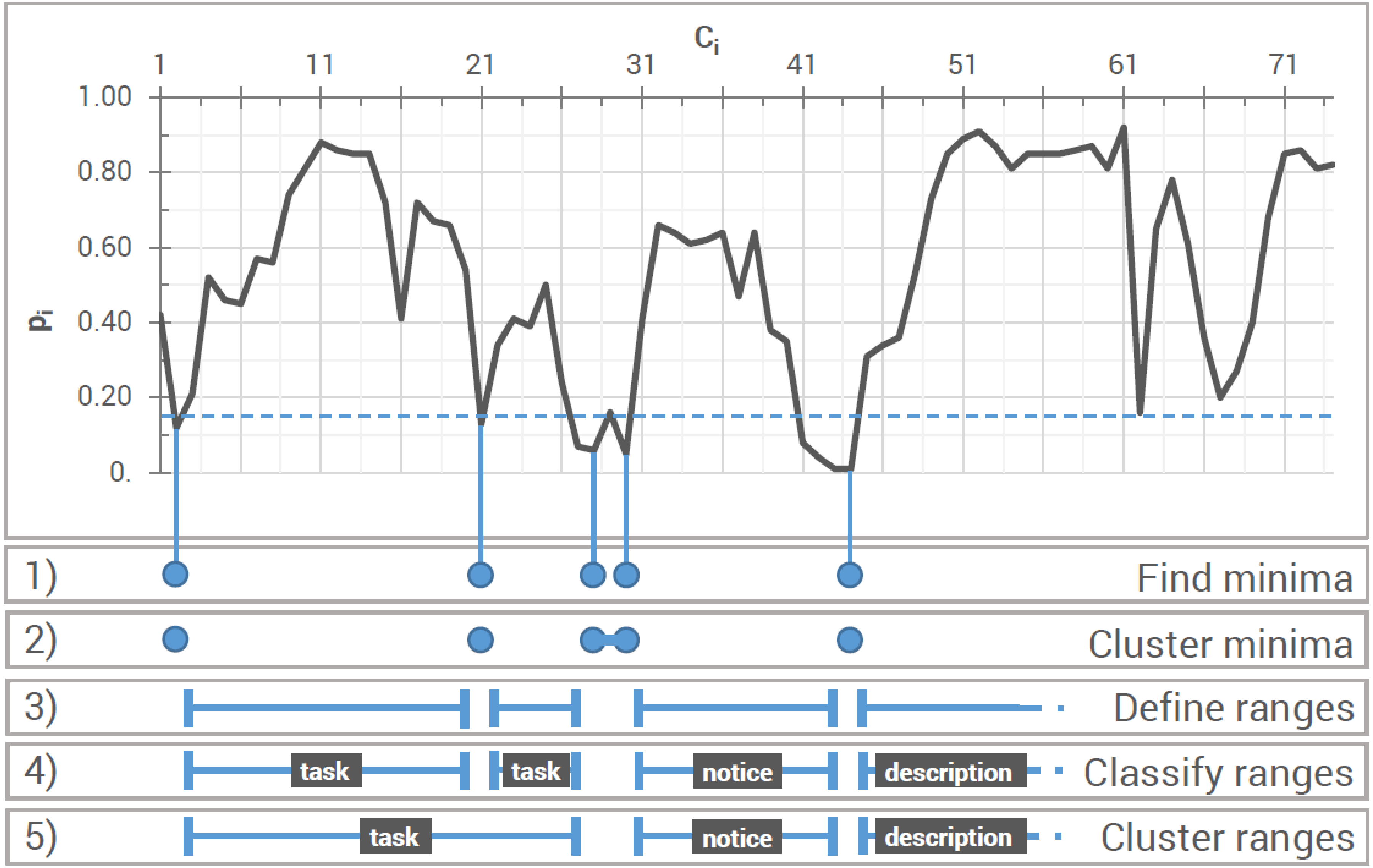
Tap Pair. Tap Done. Info: If the process was

Class: Task
Confidence: Low

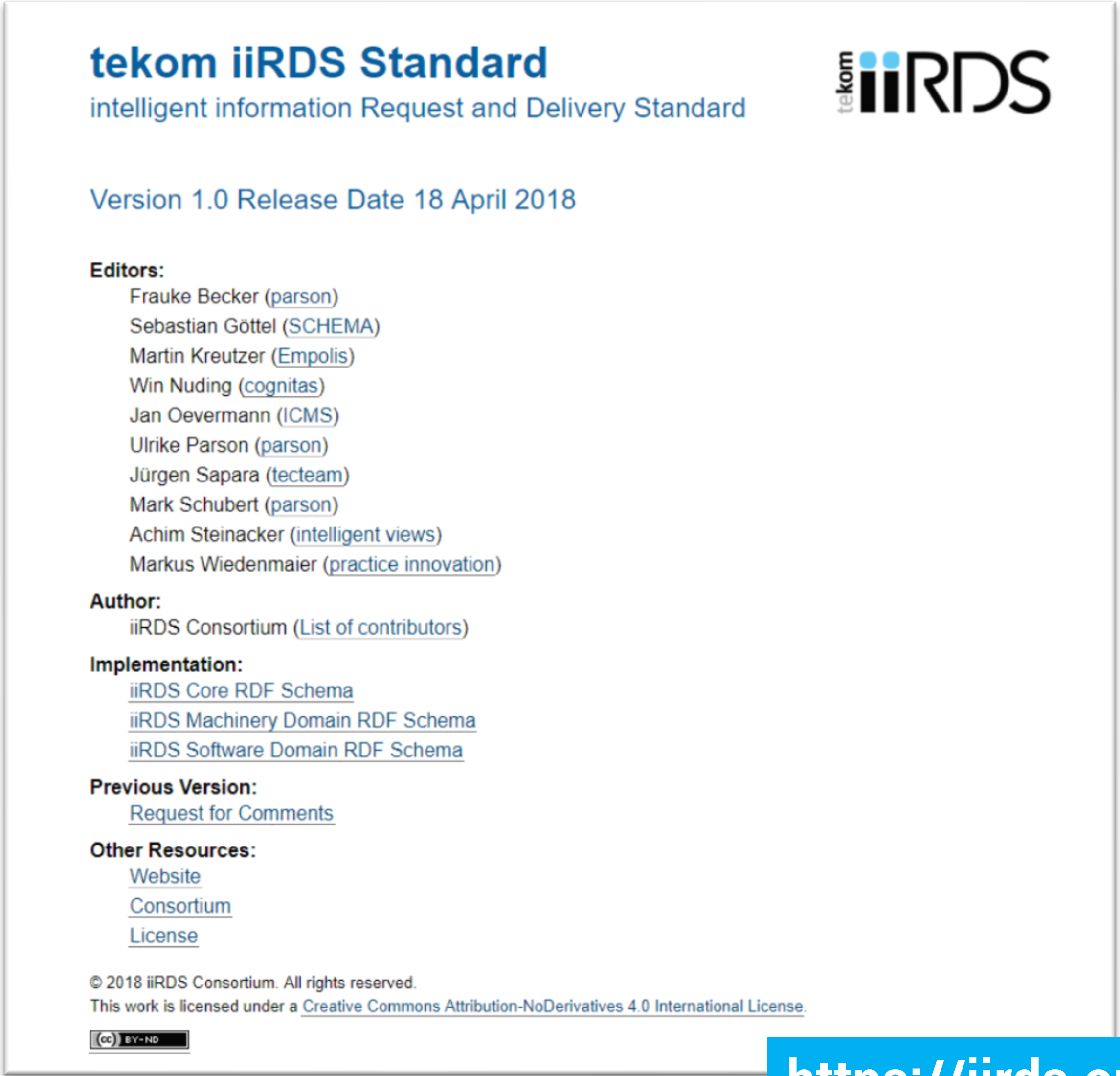
Chunking / Classification



Range finding



Metadata generation



tekom iiRDS Standard

intelligent information Request and Delivery Standard

Version 1.0 Release Date 18 April 2018

Editors:

Frauke Becker (parson)
Sebastian Göttel (SCHEMA)
Martin Kreutzer (Empolis)
Win Nuding (cognitas)
Jan Oevermann (ICMS)
Ulrike Parson (parson)
Jürgen Sapara (tecteam)
Mark Schubert (parson)
Achim Steinacker (intelligent views)
Markus Wiedenmaier (practice innovation)

Author:

iiRDS Consortium (List of contributors)

Implementation:

iiRDS Core RDF Schema
iiRDS Machinery Domain RDF Schema
iiRDS Software Domain RDF Schema

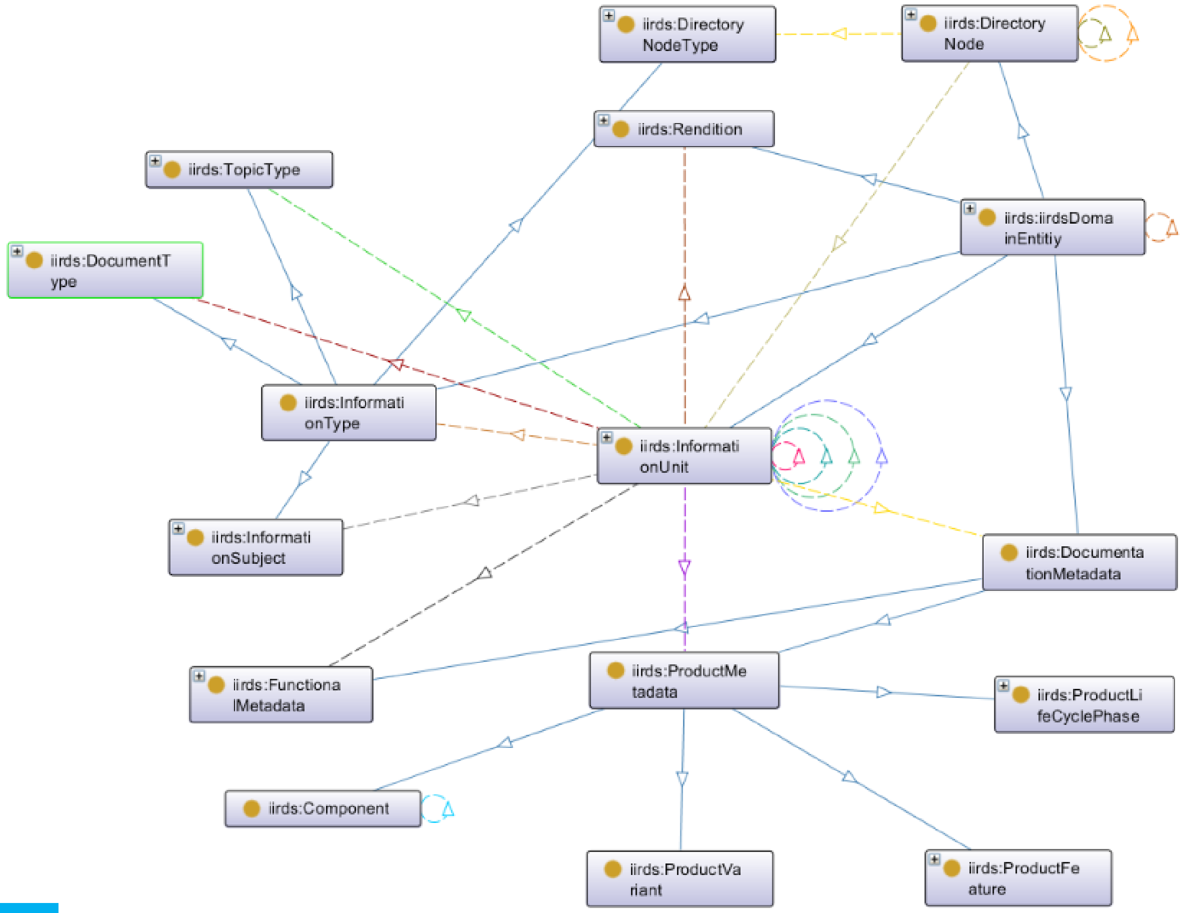
Previous Version:

Request for Comments

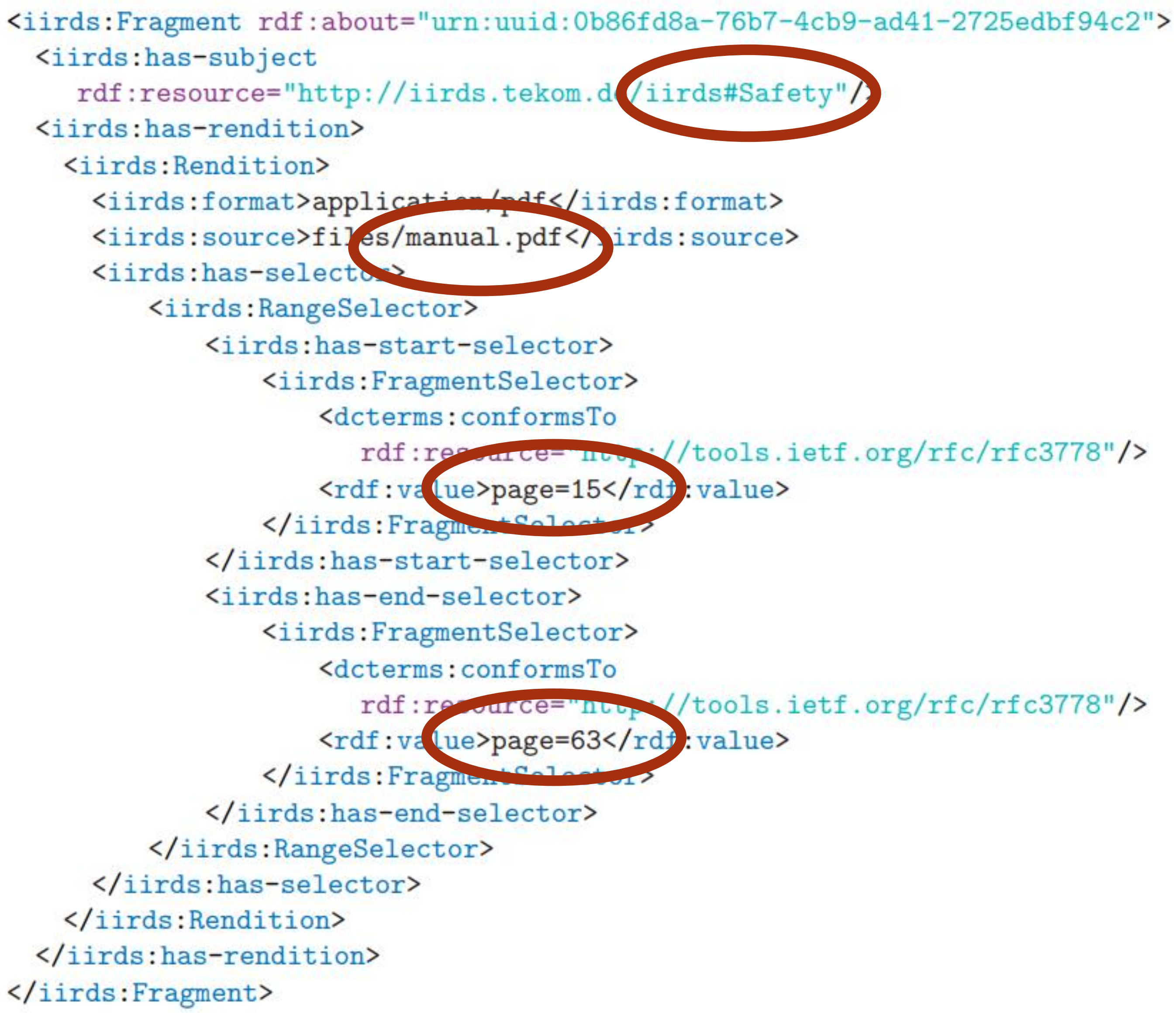
Other Resources:

Website
Consortium
License

https://iirds.org/

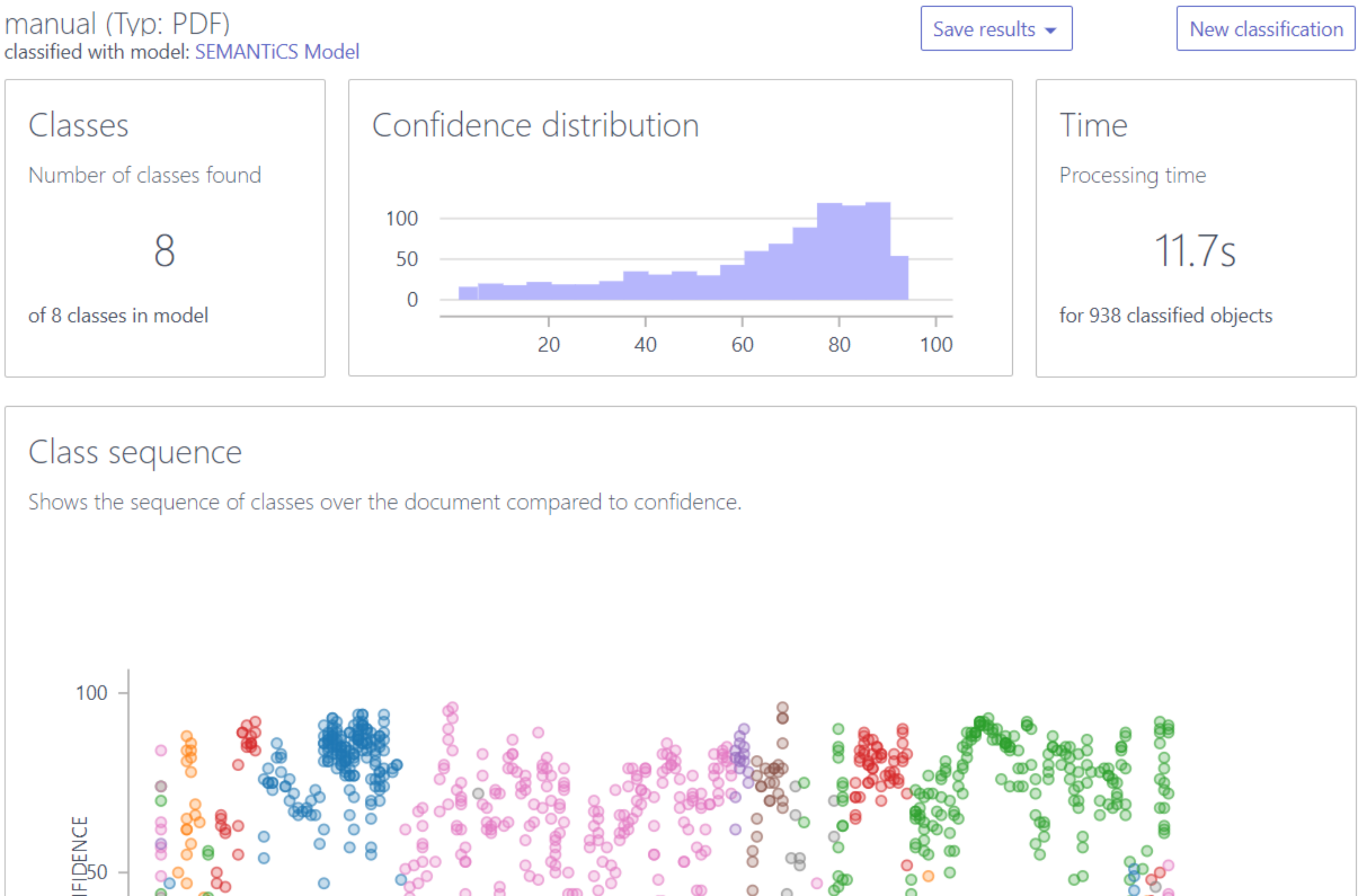


Metadata generation



```
<iirds:Fragment rdf:about="urn:uuid:0b86fd8a-76b7-4cb9-ad41-2725edbf94c2">
 <iirds:has-subject
   rdf:resource="http://iirds.tekom.de/iirds#Safety"/>
 <iirds:has-rendition>
  <iirds:Rendition>
   <iirds:format>application/pdf</iirds:format>
   <iirds:source>files/manual.pdf</iirds:source>
   <iirds:has-selector>
     <iirds:RangeSelector>
       <iirds:has-start-selector>
         <iirds:FragmentSelector>
           <dcterms:conformsTo
             rdf:resource="http://tools.ietf.org/rfc/rfc3778"/>
           <rdf:value>page=15</rdf:value>
         </iirds:FragmentSelector>
       </iirds:has-start-selector>
       <iirds:has-end-selector>
         <iirds:FragmentSelector>
           <dcterms:conformsTo
             rdf:resource="http://tools.ietf.org/rfc/rfc3778"/>
           <rdf:value>page=63</rdf:value>
         </iirds:FragmentSelector>
       </iirds:has-end-selector>
     </iirds:RangeSelector>
   </iirds:has-selector>
  </iirds:Rendition>
 </iirds:has-rendition>
</iirds:Fragment>
```

Live demo



manual (Typ: PDF)
classified with model: SEMANTiCS Model

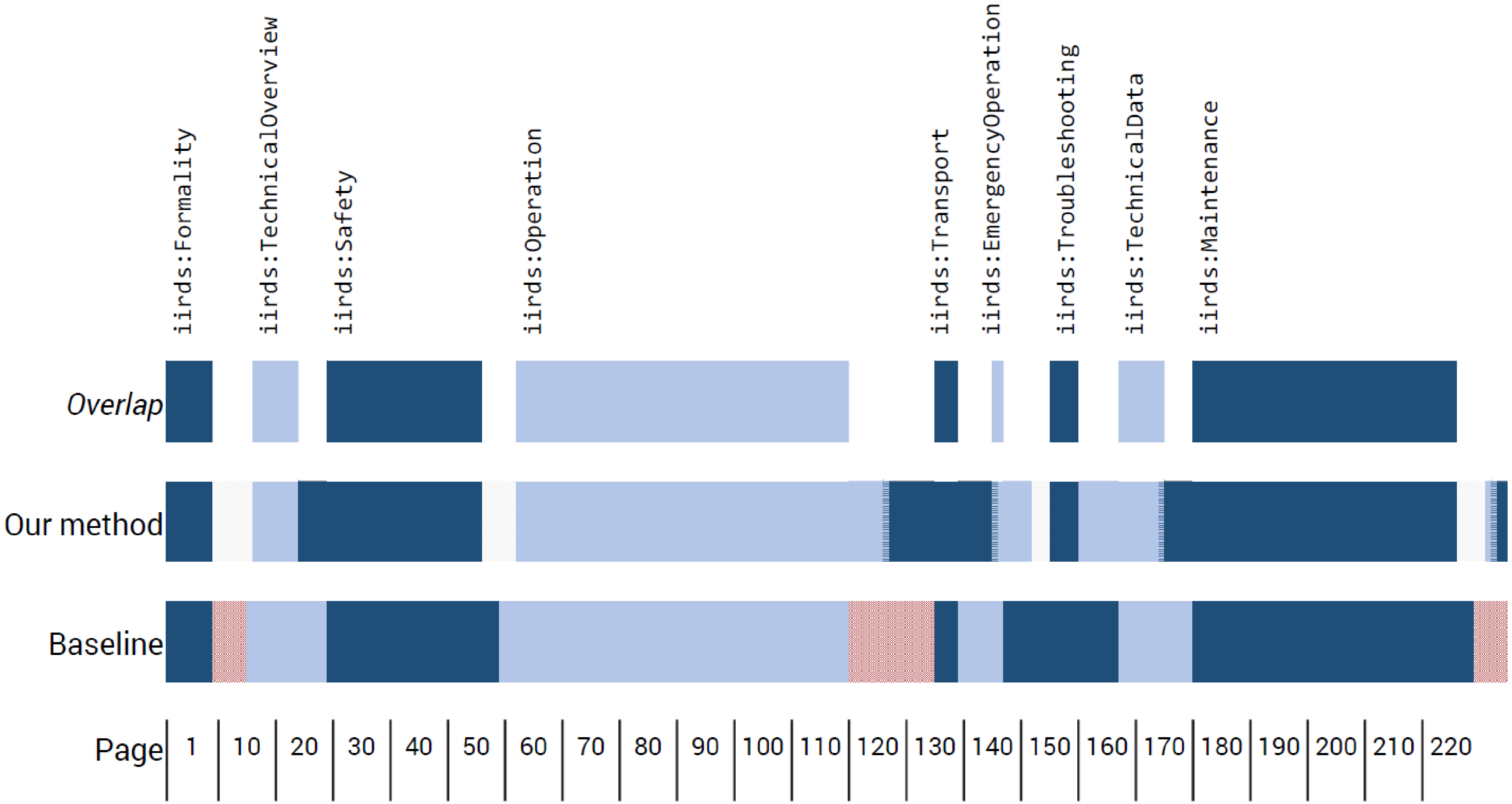
Classes
Number of classes found
8
of 8 classes in model

Confidence distribution

Time
Processing time
11.7s
for 938 classified objects

Class sequence
Shows the sequence of classes over the document compared to confidence.

Results


iirds:Formality
iirds:TechnicalOverview
iirds:Safety
iirds:Operation
iirds:Transport
iirds:EmergencyOperation
iirds:Troubleshooting
iirds:TechnicalData
iirds:Maintenance
Overlap
Our method
Baseline
Page
1
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220

Outlook & Conclusion

Outlook

- Other text sorts (e.g. patents) or document types (e.g. Word)
- Combination with other techniques (formatting / heuristics)

Conclusion

- Method relies on text and is formatting-independent
- No splitting of PDF, just additional metadata
- Good results in detecting semantic segments
- Identified ranges can be provided in a standardized format

Contact

Jan Oevermann

jan.oevermann@dfki.de

www.janoevermann.de

Code & Demo

github.com/j-oe/segments

segments.fastclass.de