**Problematic cases with Horizonte PDFs (after conversion from .tetml)**

1. Pages with map images – mostly ‘Vor Ort’ articles
   * Often have text from the maps in self-contained divs within the article and also among divs containing article content.

**potential rule:** ignore articles/pages with many div tags that are only one line. (???)

e.g.

ignore all '<div>.\*[^\n].\*</div>'

or

if num of '<div>.\*[^\n].\*</div>' > num of '<div>.\*\n.\*</div>' on page:

skip page

|  |  |  |  |
| --- | --- | --- | --- |
| **ISSUE** | **Example case from XML** | **Article tag** | **Notes** |
| 2013\_96 |  | <article id="9" page="26" title="Kampf um ein intaktes Ökosystem auf den Seychellen"> | Vor Ort |
| 2013\_97 | <div>beeindruckt, das man hier verspürt – vor­</div>  <div>treibe, veränderte sich ihre Haltung. Sie</div>  <div>ausgesetzt, man hat ein Auto.</div>  <div>akzeptierten mich, und wir konnten uns</div>  <div>Doch ich bin nicht in Kalifornien, um</div>  <div>austauschen.</div>  <div>Delfine oder Gegenwartskunst zu erfor­</div>  <div>Das persönliche Engagement geht für</div>  <div>schen. Mein Alltag ist das Street Workout,</div>  <div>mich weiter, als dass es nur die Kontakteine</div>  <div>Art urbaner Sport, der in den öffent­</div>  <div>aufnahme ermöglicht. Über die Teilnahme</div> | <article id="8" page="26" title="Street Workout – Gymnastik in Los Angeles"> | Vor Ort |
| 2013\_98 | <div>Lynda Dunkel bricht jede  Nacht um vier Uhr in den  Dschungel auf, um Orang-  Utans zu beobachten. Das  Paarungsverhalten der  Menschenaffen hängt  stärker von der Umwelt  ab als bisher vermutet.</div>  <div>THAILAND</div>  <div>Sumatra</div>  <div>VIETNAM</div>  <div>MALAYSIA</div>  <div>Borneo</div>  <div>INDONESIEN</div>  <div>auf Borneo ist der Regenwald  schon oft gerodet worden,  deshalb sind die Bäume</div>  <div>«Hier  kleiner als auf Sumatra, der</div>  <div>zweiten Heimatinsel der Orang-Utans oder  Waldmenschen, wie ihr malaiischer Name… | <article id="8" page="26" title="Mächtige Männchen auf Borneo"> | Vor Ort |
| 2013\_99 |  | <article id="8" page="26" title="Sternenkunde in Chile"> | Vor Ort |
| 2014\_101 |  | <article id="12" page="26" title="In Moskau: Die Magie der Wurzeln"> | Vor Ort |
| 2014\_102 |  | <article id="10" page="24" title="Fehl ernährung in Indien"> | Vor Ort |
| 2014\_103 | <div>45˚W 30˚W  15˚W  0˚  15˚E 30˚E 45˚E 60˚E 75˚E 90˚E  105˚E 120˚E 135˚E 150˚E 165˚E 180˚</div>  <div>für Tropen- und humanitäre Medizin A r c t i c O c e a n des</div>  <div>universitätsspitals genf, koordinierte dazu  den ersten randomisierten klinischen  Versuch in Nepal.</div> | <article id="9" page="24" title="Das Rennen gegen die Uhr"> | Vor Ort |
| 2015\_105 |  | <article id="13" page="32" title="In Sambia bei den Aquariumfischen"> | Vor Ort |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| 2018\_118 | <div>NIGERIA</div>  <div>165˚W 150˚W  135˚W  120˚W  105˚W  90˚W  75˚W  60˚W  45˚W  30˚W  15˚W  0˚  15˚E 30˚E 45˚E 60˚E 75˚E 90˚E  105˚E 120˚E 135˚E 150˚E 165˚E</div>  <div>und ihnen oft die Mittel dazu fehlen, | <article id="18" page="38" title="Krebskranke Frauen in Burkina Faso"> | Vor Ort |

1. Articles with many statitistics/graphics – mostly ‘schwerpunkt’ articles.

|  |  |  |  |
| --- | --- | --- | --- |
| **ISSUE** | **Example case from XML** | **Article tag** | **Notes** |
| 2013\_98 | <page number="17" width="623.62" height="807.87">  <div>16:00–16:30</div>  <div>15:30–16:00</div>  <div>15:00–15:30</div>  <div>14:30–15:00</div>  <div>14:00–14:30</div>  <div>Legende</div>  <div>männlich</div>  <div>weiblich Nervosität und Aufregung Zusammengehörigkeit &amp; Interaktion Alter 0–18 Alter 18+</div>  <div>Ausgewählte Beispiele</div>  <div>verärgerte Frau rügt Ehemann Vater und Sohn bezahlen, Mutter und Tochter packen ein befreundete Mütter mit Kinderwagen</div>  <page number="18" width="623.62" height="807.87">  <div>100%</div>  <div>80%</div>  <div>60%</div>  <div>40%</div>  <div>20%</div>  <div>12:00 14:00 16:00 18:00 20:00 22:00 24:00</div>  <div>0%</div>  <div>18 Verhalten Schweizerischer am Fussgängerstreifen zwischen Nationalfonds Neuengasse und Bahnhofplatz Bern, – Werktag Akademien 2009 Schweiz: 31% 69% Horizonte Nr. 98</div>  </page> | <article id="5" page="10" title="Denken in Zahlen"> | Schwerpunkt |
| 2014\_101 | <div>Schwerpunkt Internationalität der Forschung</div>  <div>mobilität</div>  <div>Ausländeranteil unter den Forschenden in % (2011)</div>  <div>– davon aus welchem Land (%; nur Anteile über 10%)</div>  <div>Wiederkehrrate der Forschenden mit  Auslanderfahrung (%)</div>  <div>Schweiz</div>  <div>Kanada  Australien  USA  Schweden  Vereinigtes Königreich  Niederlande  Deutschland</div>  <div>Dänemark</div>  <div>Belgien  Frankreich</div>  <div>Spanien  Brasilien</div>  <div>Japan  Italien  Indien</div> | <article id="10" page="20" title="Die internationale Verflechtung des Schweizer Forschungsplatzes"> | Schwerpunkt |
| 2016\_109 | <div>Und 2014 wurde die Bestimmung aufgeweicht,  wonach kein Kriegsmaterial an Staaten geliefert  wird, die systematisch und schwerwiegend die  Menschenrechte verletzen: Solche Exporte sind  neu erlaubt, wenn das Risiko, dass die Menschenrechtsverletzungen  mit dem Schweizer  Material geschehen, gering ist.</div>  <div>Belgien</div>  <div>Kanada</div>  <div>Frankreich</div>  <div>Schweiz</div>  <div>26.3</div>  <div>53.2</div>  <div>24.4</div>  <div>USA</div>  <div>Spanien</div>  <div>Italien</div>  <div>Marokko</div>  <div>Mexiko</div>  <div>Algerien</div>  <div>Kolumbien</div>  <div>Bolivien</div>  <div>Brasilien</div>  <div>Chile</div>  <div>Argentinien</div>  <div>tschechische  Republik  Österreich  Slowenien</div>  <div>Marcel Hänggi ist Wissenschaftsjournalist in Zürich.</div>  <div>18 Schweizerischer Nationalfonds – Akademien Schweiz: Horizonte Nr. 109</div> | <article id="8" page="18" title="Umstrittene Exportware"> | Schwerpunkt |

1. Miscellaneous

|  |  |  |  |
| --- | --- | --- | --- |
| **ISSUE** | **Example case from XML** | **Article/parent tag** | **Notes** |
| 2013\_97 | <div>M Vergils Akrostichon</div>  <div>anch eine wird sich dunkel  <div>C. Castelletti, Following Aratus’ plow: Vergil’s  signature in the Aeneid, Museum Helveticum  2012: 69, 83–95.</div>  <div>ikimedia Commons</div>  <div>W</div>  <div>Omnipotentes Medium: In der aktuellen Fantasyliteratur  spielen Zauberbücher eine wichtige  Rolle.</div> | <article id="20" page="45" title="Wie das Zauberbuch das Buch verzaubert «Vom Dräckloch i Himel» Vergils Akrostichon"> |  |
| 2013\_98 | <div>versus «quantitativ»  – dieser Dämon verfolgt  mich seit vielen Jahren.</div>  <div>«Qualitativ» | <article id="4" page="8" title="Interpretieren oder rechnen? Qualitative versus quantitative Sozialforschung">  …  <page number="9" width="623.62" height="807.87"> | Konferenz |
| 2015\_104 | <div>Und die eigenständige Forschung?  Rosetta hat bei allem das letzte Wort.  Meine gesamte Kommunikation zur Erde  läuft über ihre Signalverstärker. «Rosetta»  hat in Google 46 Millionen Treffer. Die  Süddeutsche Zeitung schreibt über meine  Forschungstätigkeit: «Rosetta selbst  ist deutlich produktiver.» Und natürlich  hatte Rosetta gerade wieder eine ganze  Reihe Science-Papers, ohne mich</div>  <div>»</div>  <div>darin auch nur zu erwähnen.</div>  <div>Mir reicht’s. Ich lege ein Bein  hoch und schalte ab.</div> | <article id="9" page="22" title="Nicht sehr sonnig"> |  |
| 2014\_103 | This process depends on considerable interaction with researchers who think differently, who criticise the original idea, who enrich it and confront', 'Dear Horizons, I am almost always pleasantly surprised by your choice of focus topics. Time and again I’m delighted by the quality of the articles, by the excellent layout and often also by the expressiveness of the professional portrait photos. And when compared to its German equivalent, the magazine Forschung of the Deutsche Forschungsgemeinschaft, Horizons also comes off considerably better, to my taste. Instead of pandering all too frequently to the vanity of academic bureaucrats, your focus is on your readership, and you offer food for thought in clear prose. There’s just one thing: on page 48 of edition No. 102 you carried a photo of a rodent that the caption claimed was a muskrat. But the photo is in fact of a nutria. I suggest that when using third-party it with their own approaches. It is this part of the scientific development process that is profoundly international. Only a confrontation with other milieus can lead science from local to universal validity. | article id="25" page="50" title="Only international scholarship can achieve universal validity">  <page number="50" width="623.62" height="807.87">  <div>Access all areas</div> | Access all areas with letter to the editor  Text is extracted vertically, and columns of each ‘article’ are not isolated. |

1. **Contents listing**

|  |  |  |  |
| --- | --- | --- | --- |
| **ISSUE** | **Example case from XML/Desrciption** | **Article/parent tag** | **Notes** |
| 69\_de / 69\_fr |  |  |  |
| 77\_fr | Article *Boite â toils* is not listed in contents even though it exists on page 26 of pdf.  Noticed only when checking the number of articles in parallel German and French issues |  | Fixed with ‘anything missing’ catchall option in contents\_control.py |
| 85\_de / 85\_fr | Offset required as front page is missing from pdf. This means all page numbers listed on contents page need to be subtracted by 1. 3 different page numbers and article titles needed to be added manually. |  | Fixed with ‘anything missing’ catchall option in contents\_control.py |
|  |  |  |  |

Mixed paragraphs in German:

2015\_105: article 6 - Algorithmen treffen auf Papyrus – fixed merging so with new script (Dec 31)

Existing problem: first paragraph is not actually first para. Maybe could be fixed using drop cap info.

<article title="Archäologische Feldforschung mit dem Tablet" id="a6" potentialErrors="False">

<div>Algorithmen treffen auf Papyrus</div>

<div>Historiker und Archäologen machen sich mit den Instrumenten der Informatik vertraut. Diese Entwicklung ist unumgänglich – und hat das Zeug, die Disziplinen neu zu erfinden. Von Fabien Goubet</div>

<div>diese Informationen mit geografischen Gegeben heiten wie etwa der Topografie des Geländes in Verbindung. Aus diesen Angaben errechnet ein Algorithmus die Wahrscheinlichkeit einer Grenzstelle zu einem bestimmten Zeitpunkt.</div>

<div>Diese digitalen Hilfsmittel ersetzen jedoch die archäologische Feldarbeit nicht, betont der Forscher. Kann dank des Modells eine Voraussage gemacht werden, setzt Sylvian Fachard seinen Archäologenhut auf und sucht vor Ort Spuren, um den angenommenen Grenzverlauf der Gemeinde zu bestätigen oder zu widerlegen. «Es ist nötig, alle Angaben zu interpretieren, um zu einem treffenden Ergebnis zu gelangen. Denn das Digitale generiert Angaben, die wir wiederum miteinander verbinden müssen. Dies erfordert ein komplexeres Analysenniveau als bisher.»</div>