SDR Swiss Disability Research

zh aw School of Engineering

Universität Zürich[UZH]

Let’s Make PDFs Accessible for All: A Survey on PDF-Accessibility in Switzerland

Prof. Dr.
Alireza Darvishy

Prof. Dr.
Rolf Sethe

Ines
Engler

Felix
Schmitt-Koopmann

Oriane
Pierrès

Agenda

Part I: Introduction

- What makes a PDF document accessible?
- The problem with accessible PDFs
- PDF Tags
- Creating a PDF UA
- PDF remediation with AI

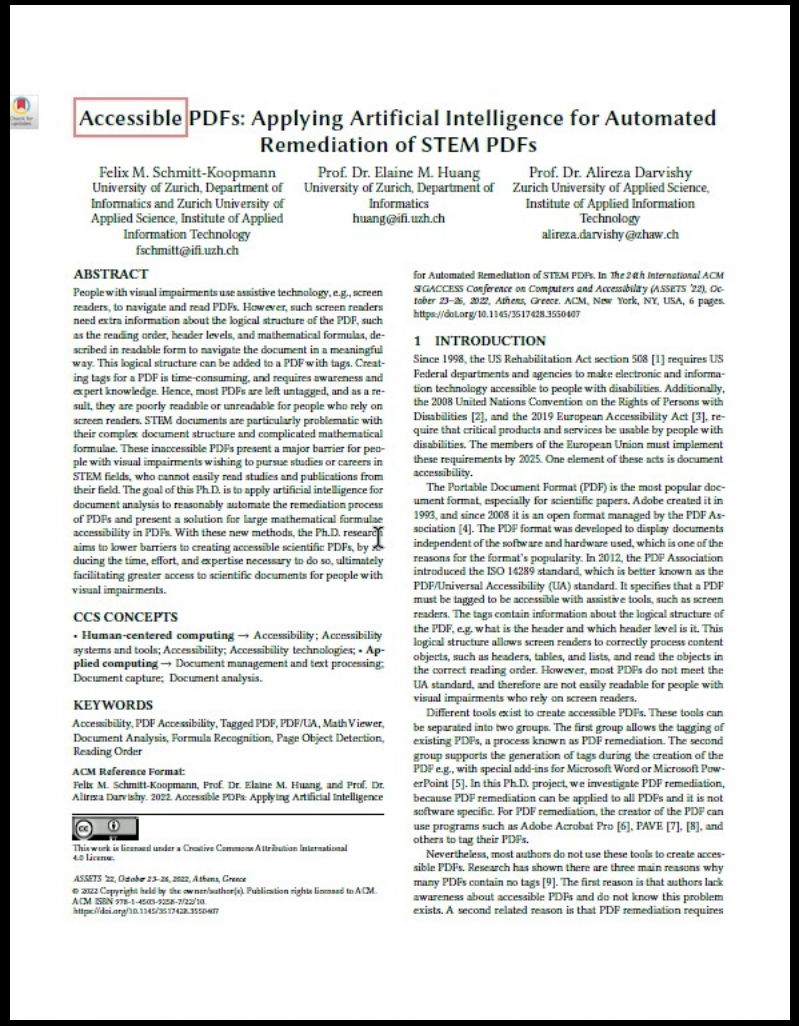
Part II: The state of accessible scientific PDFs in Switzerland

Before we begin…

What makes a PDF document accessible?

Introduction

Example: Reading Order



Accessible PDFs: Applying Artificial Intelligence for Automated Remediation of STEM PDFs

Felix M. Schmitt-Koopmann
University of Zurich, Department of Informatics and Zurich University of Applied Science, Institute of Applied Information Technology
fschmitt@ifi.uzh.ch

Prof. Dr. Elaine M. Huang
University of Zurich, Department of Informatics
huang@ifi.uzh.ch

Prof. Dr. Alireza Darvishy
Zurich University of Applied Science, Institute of Applied Information Technology
alireza.darvishy@zhaw.ch

ABSTRACT

People with visual impairments use assistive technology, e.g., screen readers, to navigate and read PDFs. However, such screen readers need extra information about the logical structure of the PDF, such as the reading order, header levels, and mathematical formulas, described in readable form to navigate the document in a meaningful way. This logical structure can be added to a PDF with tags. Creating tags for a PDF is time-consuming, and requires awareness and expert knowledge. Hence, most PDFs are left untagged, and as a result, they are poorly readable or unreadable for people who rely on screen readers. STEM documents are particularly problematic with their complex document structure and complicated mathematical formulae. These inaccessible PDFs present a major barrier for people with visual impairments wishing to pursue studies or careers in STEM fields, who cannot easily read studies and publications from their field. The goal of this Ph.D. is to apply artificial intelligence for document analysis to reasonably automate the remediation process of PDFs and present a solution for large mathematical formulae accessibility in PDFs. With these new methods, the Ph.D. research aims to lower barriers to creating accessible scientific PDFs, by reducing the time, effort, and expertise necessary to do so, ultimately facilitating greater access to scientific documents for people with visual impairments.

CCS CONCEPTS

• **Human-centered computing** → Accessibility; Accessibility systems and tools; Accessibility; Accessibility technologies; • **Applied computing** → Document management and text processing; Document capture; Document analysis.

KEYWORDS

Accessibility, PDF Accessibility, Tagged PDF, PDF/UA, Math Viewer, Document Analysis, Formula Recognition, Page Object Detection, Reading Order

ACM Reference Format:

Felix M. Schmitt-Koopmann, Prof. Dr. Elaine M. Huang, and Prof. Dr. Alireza Darvishy. 2022. Accessible PDFs: Applying Artificial Intelligence for Automated Remediation of STEM PDFs. In *The 24th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '22), October 23–26, 2022, Athens, Greece.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3517428.3550407

ASSETS '22, October 23–26, 2022, Athens, Greece

ACM ISBN 978-1-4503-9258-7/22/10
https://doi.org/10.1145/3517428.3550407

1 INTRODUCTION

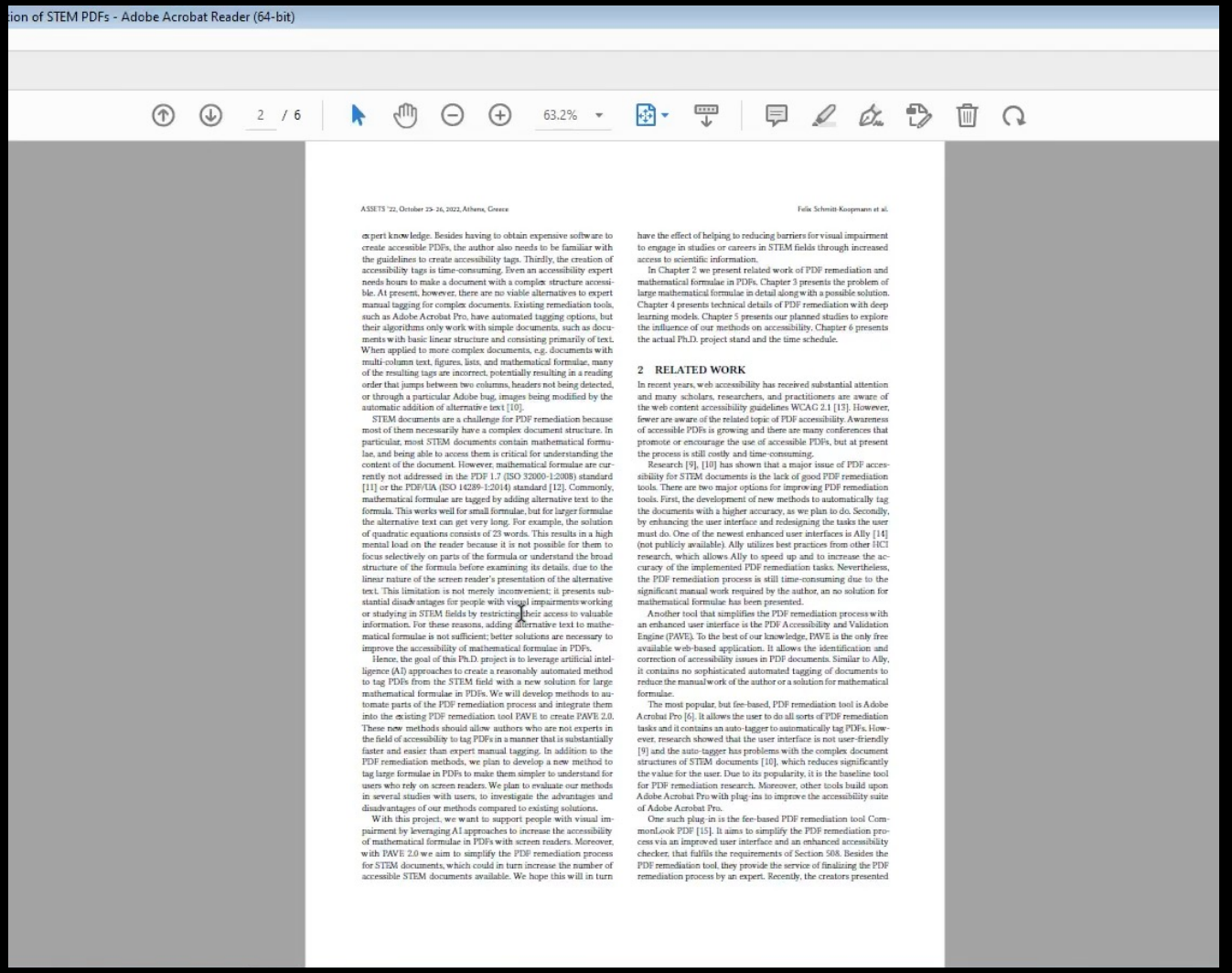
Since 1998, the US Rehabilitation Act section 508 [1] requires US Federal departments and agencies to make electronic and information technology accessible to people with disabilities. Additionally, the 2008 United Nations Convention on the Rights of Persons with Disabilities [2], and the 2019 European Accessibility Act [3], require that critical products and services be usable by people with disabilities. The members of the European Union must implement these requirements by 2025. One element of these acts is document accessibility.

The Portable Document Format (PDF) is the most popular document format, especially for scientific papers. Adobe created it in 1993, and since 2008 it is an open format managed by the PDF Association [4]. The PDF format was developed to display documents independent of the software and hardware used, which is one of the reasons for the format's popularity. In 2012, the PDF Association introduced the ISO 14289 standard, which is better known as the PDF/Universal Accessibility (UA) standard. It specifies that a PDF must be tagged to be accessible with assistive tools, such as screen readers. The tags contain information about the logical structure of the PDF, e.g. what is the header and which header level is it. This logical structure allows screen readers to correctly process content objects, such as headers, tables, and lists, and read the objects in the correct reading order. However, most PDFs do not meet the UA standard, and therefore are not easily readable for people with visual impairments who rely on screen readers.

Different tools exist to create accessible PDFs. These tools can be separated into two groups. The first group allows the tagging of existing PDFs, a process known as PDF remediation. The second group supports the generation of tags during the creation of the PDF e.g., with special add-ins for Microsoft Word or Microsoft PowerPoint [5]. In this Ph.D. project, we investigate PDF remediation, because PDF remediation can be applied to all PDFs and it is not software specific. For PDF remediation, the creator of the PDF can use programs such as Adobe Acrobat Pro [6], PAVE [7], [8], and others to tag their PDFs.

Nevertheless, most authors do not use these tools to create accessible PDFs. Research has shown there are three main reasons why many PDFs contain no tags [9]. The first reason is that authors lack awareness about accessible PDFs and do not know this problem exists. A second related reason is that PDF remediation requires

Example: Headings



Example: Visuals

P8 – Make visuals accessible

Example: Tables

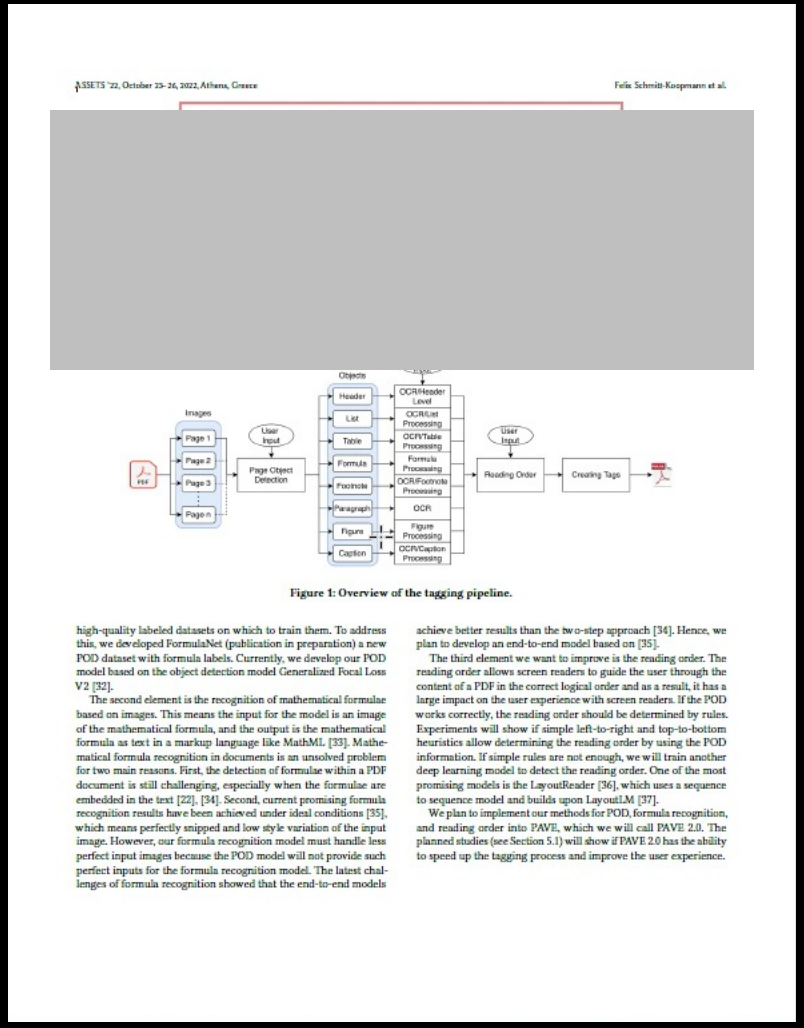


Figure 1: Overview of the tagging pipeline.

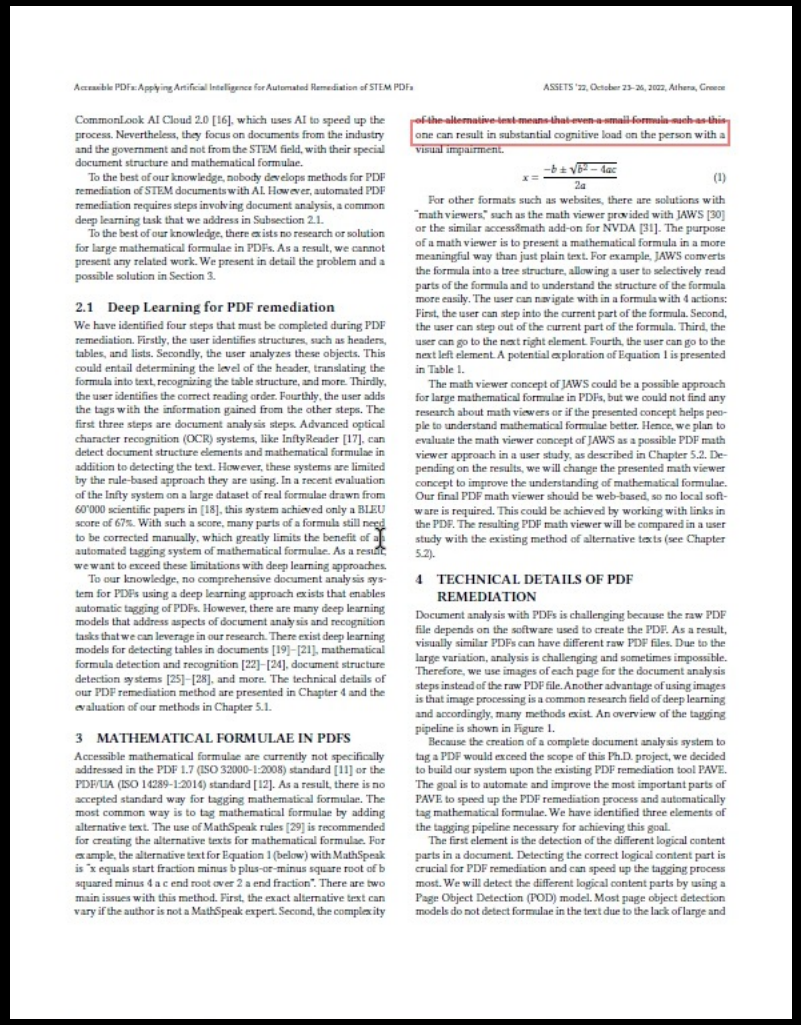
high-quality labeled datasets on which to train them. To address this, we developed FormulaNet (publication in preparation) a new POD dataset with formula labels. Currently, we develop our POD model based on the object detection model Generalized Focal Loss V2 [32].

The second element is the recognition of mathematical formulae based on images. This means the input for the model is an image of the mathematical formula, and the output is the mathematical formula as text in a markup language like MathML [33]. Mathematical formula recognition in documents is an unsolved problem for two main reasons. First, the detection of formulae within a PDF document is still challenging, especially when the formulae are embedded in the text [22], [34]. Second, current promising formula recognition results have been achieved under ideal conditions [35], which means perfectly snipped and low style variation of the input image. However, our formula recognition model must handle less perfect input images because the POD model will not provide such perfect inputs for the formula recognition model. The latest challenges of formula recognition showed that the end-to-end models achieve better results than the two-step approach [34]. Hence, we plan to develop an end-to-end model based on [35].

The third element we want to improve is the reading order. The reading order allows screen readers to guide the user through the content of a PDF in the correct logical order and as a result, it has a large impact on the user experience with screen readers. If the POD works correctly, the reading order should be determined by rules. Experiments will show if simple left-to-right and top-to-bottom heuristics allow determining the reading order by using the POD information. If simple rules are not enough, we will train another deep learning model to detect the reading order. One of the most promising models is the LayoutReader [36], which uses a sequence to sequence model and builds upon LayoutLM [37].

We plan to implement our methods for POD, formula recognition, and reading order into PAVE, which we will call PAVE 2.0. The planned studies (see Section 5.1) will show if PAVE 2.0 has the ability to speed up the tagging process and improve the user experience.

Example: Mathematical Formulas



Accessible PDFs: Applying Artificial Intelligence for Automated Remediation of STEM PDFs ASSETS '22, October 23–26, 2022, Athens, Greece

CommonLook AI Cloud 2.0 [16], which uses AI to speed up the process. Nevertheless, they focus on documents from the industry and the government and not from the STEM field, with their special document structure and mathematical formulae.

To the best of our knowledge, nobody develops methods for PDF remediation of STEM documents with AI. However, automated PDF remediation requires steps involving document analysis, a common deep learning task that we address in Subsection 2.1.

To the best of our knowledge, there exists no research or solution for large mathematical formulae in PDFs. As a result, we cannot present any related work. We present in detail the problem and a possible solution in Section 3.

2.1 Deep Learning for PDF remediation

We have identified four steps that must be completed during PDF remediation. Firstly, the user identifies structures, such as headers, tables, and lists. Secondly, the user analyzes these objects. This could entail determining the level of the header, translating the formula into text, recognizing the table structure, and more. Thirdly, the user identifies the correct reading order. Fourthly, the user adds the tags with the information gained from the other steps. The first three steps are document analysis steps. Advanced optical character recognition (OCR) systems, like InftyReader [17], can detect document structure elements and mathematical formulae in addition to detecting the text. However, these systems are limited by the rule-based approach they are using. In a recent evaluation of the Infty system on a large dataset of real formulae drawn from 60'000 scientific papers in [18], this system achieved only a BLEU score of 67%. With such a score, many parts of a formula still need to be corrected manually, which greatly limits the benefit of an automated tagging system of mathematical formulae. As a result, we want to exceed these limitations with deep learning approaches.

To our knowledge, no comprehensive document analysis system for PDFs using a deep learning approach exists that enables automatic tagging of PDFs. However, there are many deep learning models that address aspects of document analysis and recognition tasks that we can leverage in our research. There exist deep learning models for detecting tables in documents [19]–[21], mathematical formula detection and recognition [22]–[24], document structure detection systems [25]–[28], and more. The technical details of our PDF remediation method are presented in Chapter 4 and the evaluation of our methods in Chapter 5.1.

3 MATHEMATICAL FORMULAE IN PDFS

Accessible mathematical formulae are currently not specifically addressed in the PDF 1.7 (ISO 32000-1:2008) standard [11] or the PDF/UA (ISO 14289-1:2014) standard [12]. As a result, there is no accepted standard way for tagging mathematical formulae. The most common way is to tag mathematical formulae by adding alternative text. The use of MathSpeak rules [29] is recommended for creating the alternative texts for mathematical formulae. For example, the alternative text for Equation 1 (below) with MathSpeak is "x equals start fraction minus b plus-or-minus square root of b squared minus 4 a c end root over 2 a end fraction". There are two main issues with this method. First, the exact alternative text can vary if the author is not a MathSpeak expert. Second, the complexity of the alternative text means that even a small formula such as this one can result in substantial cognitive load on the person with a visual impairment.

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (1)$$

For other formats such as websites, there are solutions with "math viewers," such as the math viewer provided with JAWS [30] or the similar access8math add-on for NVDA [31]. The purpose of a math viewer is to present a mathematical formula in a more meaningful way than just plain text. For example, JAWS converts the formula into a tree structure, allowing a user to selectively read parts of the formula and to understand the structure of the formula more easily. The user can navigate with in a formula with 4 actions: First, the user can step into the current part of the formula. Second, the user can step out of the current part of the formula. Third, the user can go to the next right element. Fourth, the user can go to the next left element. A potential exploration of Equation 1 is presented in Table 1.

The math viewer concept of JAWS could be a possible approach for large mathematical formulae in PDFs, but we could not find any research about math viewers or if the presented concept helps people to understand mathematical formulae better. Hence, we plan to evaluate the math viewer concept of JAWS as a possible PDF math viewer approach in a user study, as described in Chapter 5.2. Depending on the results, we will change the presented math viewer concept to improve the understanding of mathematical formulae. Our final PDF math viewer should be web-based, so no local software is required. This could be achieved by working with links in the PDF. The resulting PDF math viewer will be compared in a user study with the existing method of alternative texts (see Chapter 5.2).

4 TECHNICAL DETAILS OF PDF REMEDIATION

Document analysis with PDFs is challenging because the raw PDF file depends on the software used to create the PDF. As a result, visually similar PDFs can have different raw PDF files. Due to the large variation, analysis is challenging and sometimes impossible. Therefore, we use images of each page for the document analysis steps instead of the raw PDF file. Another advantage of using images is that image processing is a common research field of deep learning and accordingly, many methods exist. An overview of the tagging pipeline is shown in Figure 1.

Because the creation of a complete document analysis system to tag a PDF would exceed the scope of this Ph.D. project, we decided to build our system upon the existing PDF remediation tool PAVE. The goal is to automate and improve the most important parts of PAVE to speed up the PDF remediation process and automatically tag mathematical formulae. We have identified three elements of the tagging pipeline necessary for achieving this goal.

The first element is the detection of the different logical content parts in a document. Detecting the correct logical content part is crucial for PDF remediation and can speed up the tagging process most. We will detect the different logical content parts by using a Page Object Detection (POD) model. Most page object detection models do not detect formulae in the text due to the lack of large and

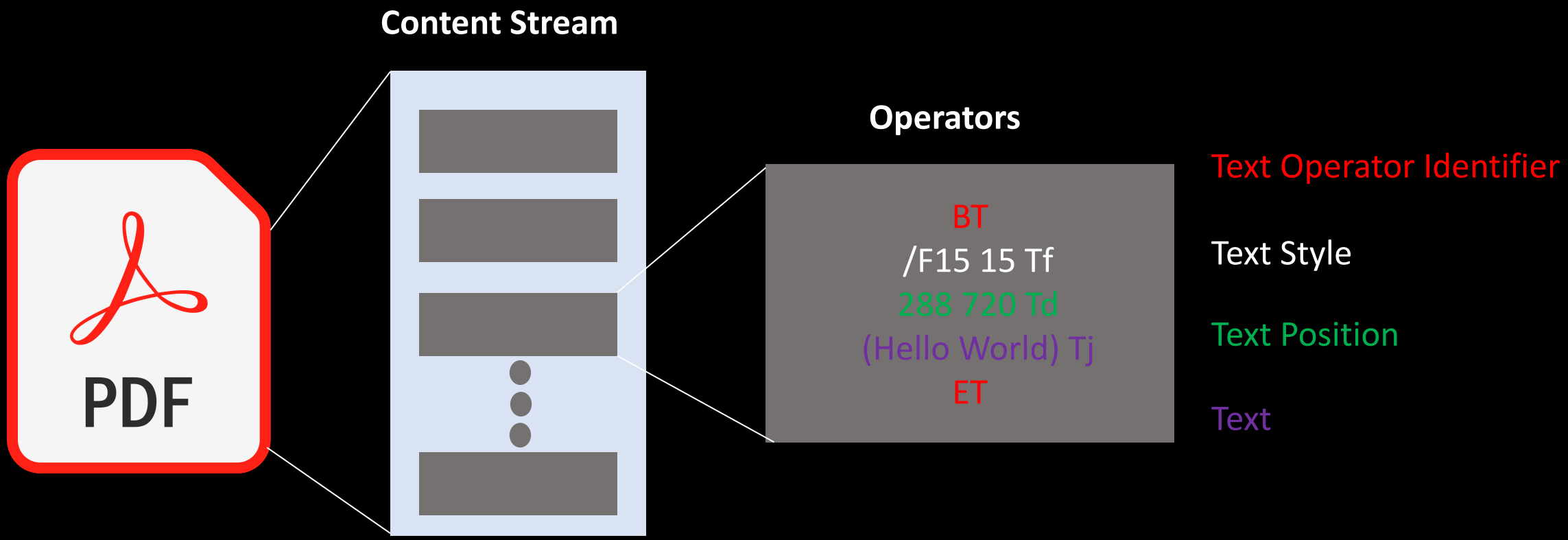
one can result in substantial cognitive load on the person with a visual impairment.

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (1)$$

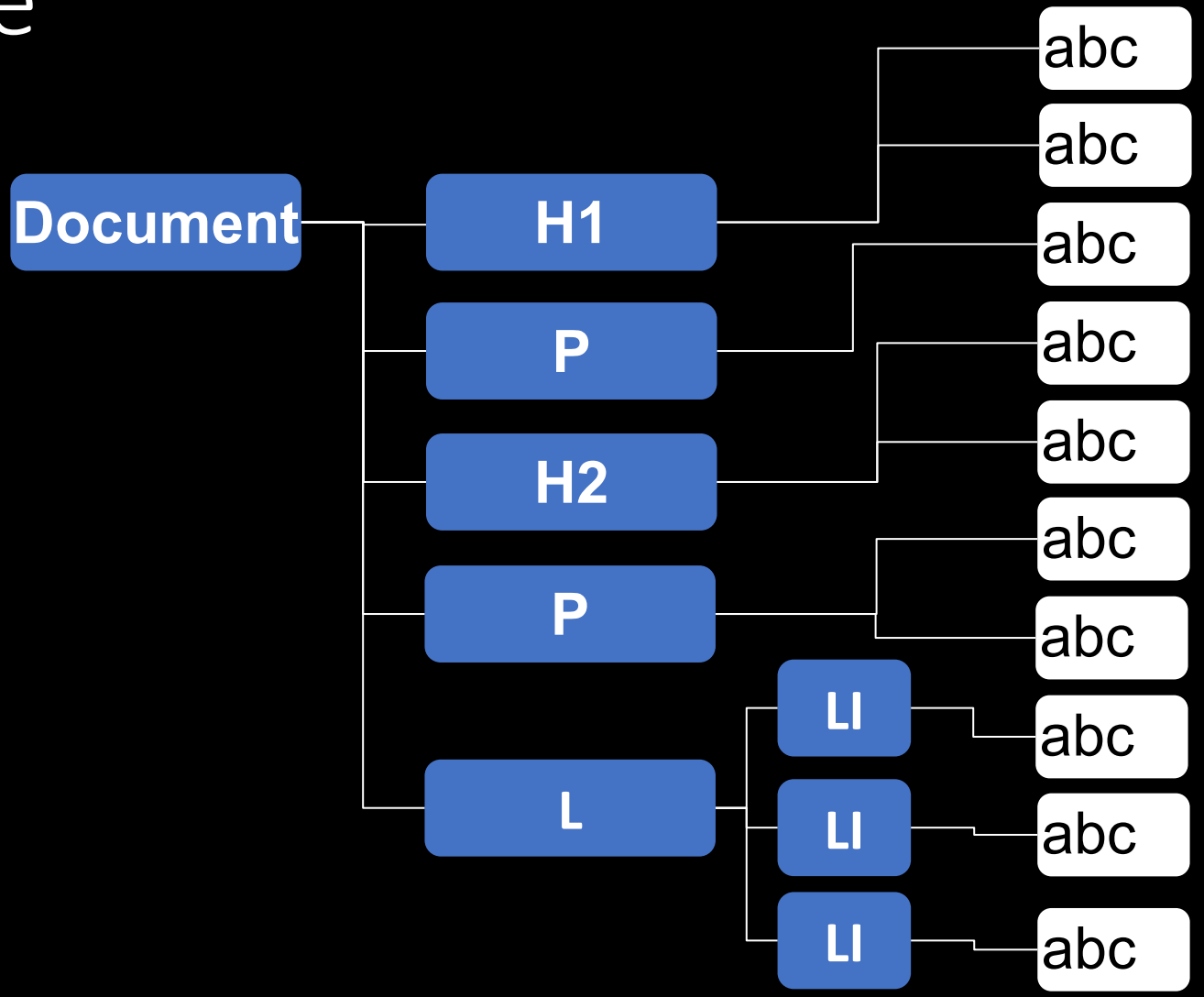
For other formats such as websites, there are solutions with "math viewers," such as the math viewer provided with JAWS [30]

What have you noticed about the slides?

Why is the PDF format problematic?



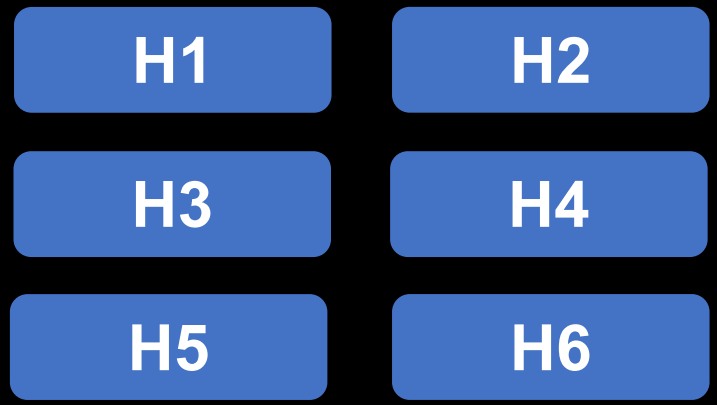
Struct Tree


Document
H1
P
H2
P
L
LI
LI
LI
abc
abc
abc
abc
abc
abc
abc
abc
abc
abc

Heading – Hn tags

Heading 1 (H1)
- Heading 1.1 (H2)
  - Heading 1.1.1 (H3)
- Heading 1.2 (H2)

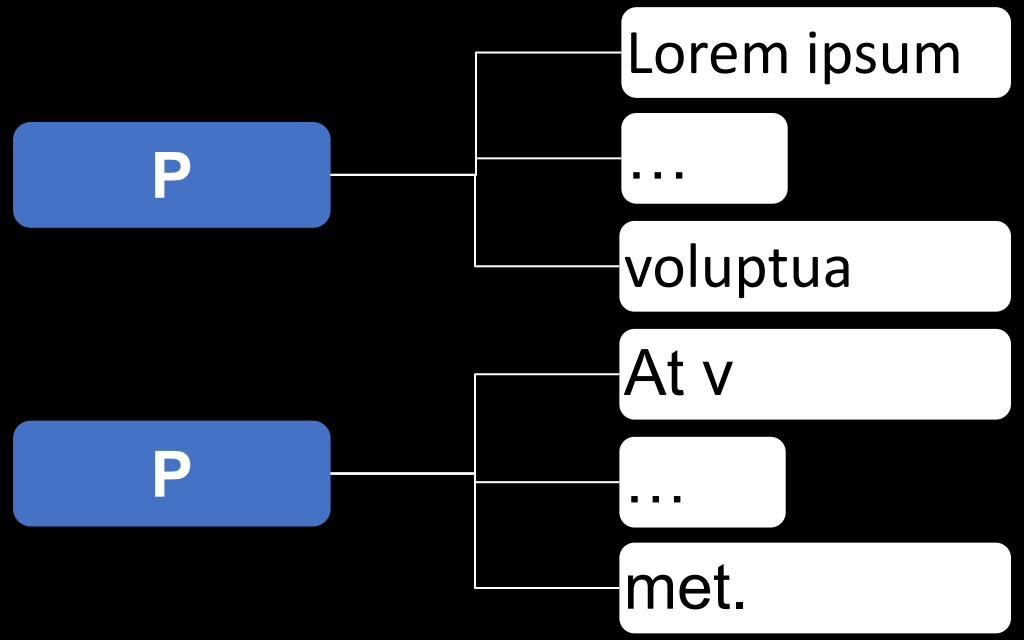
Heading 2 (H1)



Paragraph – P tags

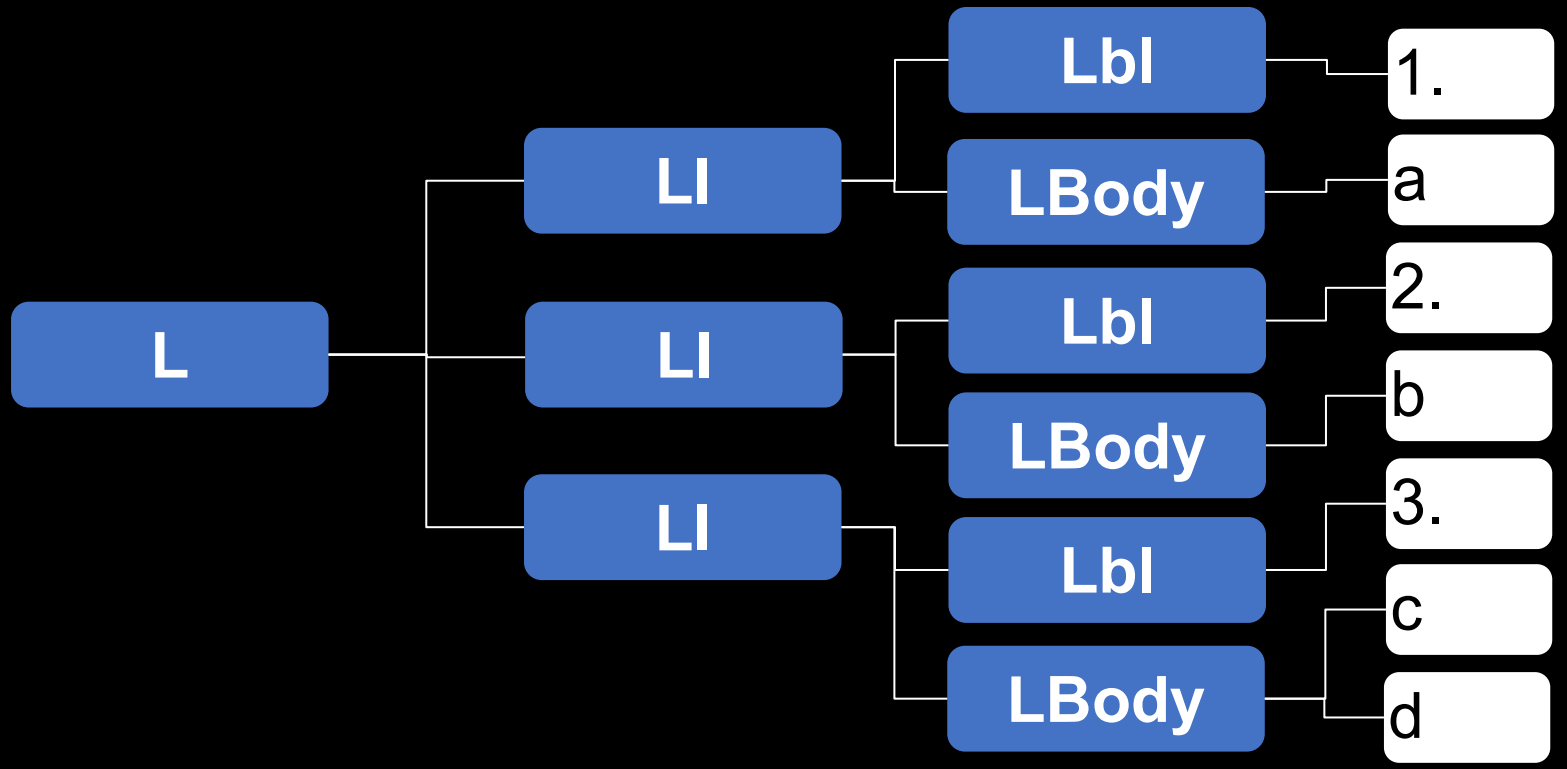
(P) Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

(P) At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.



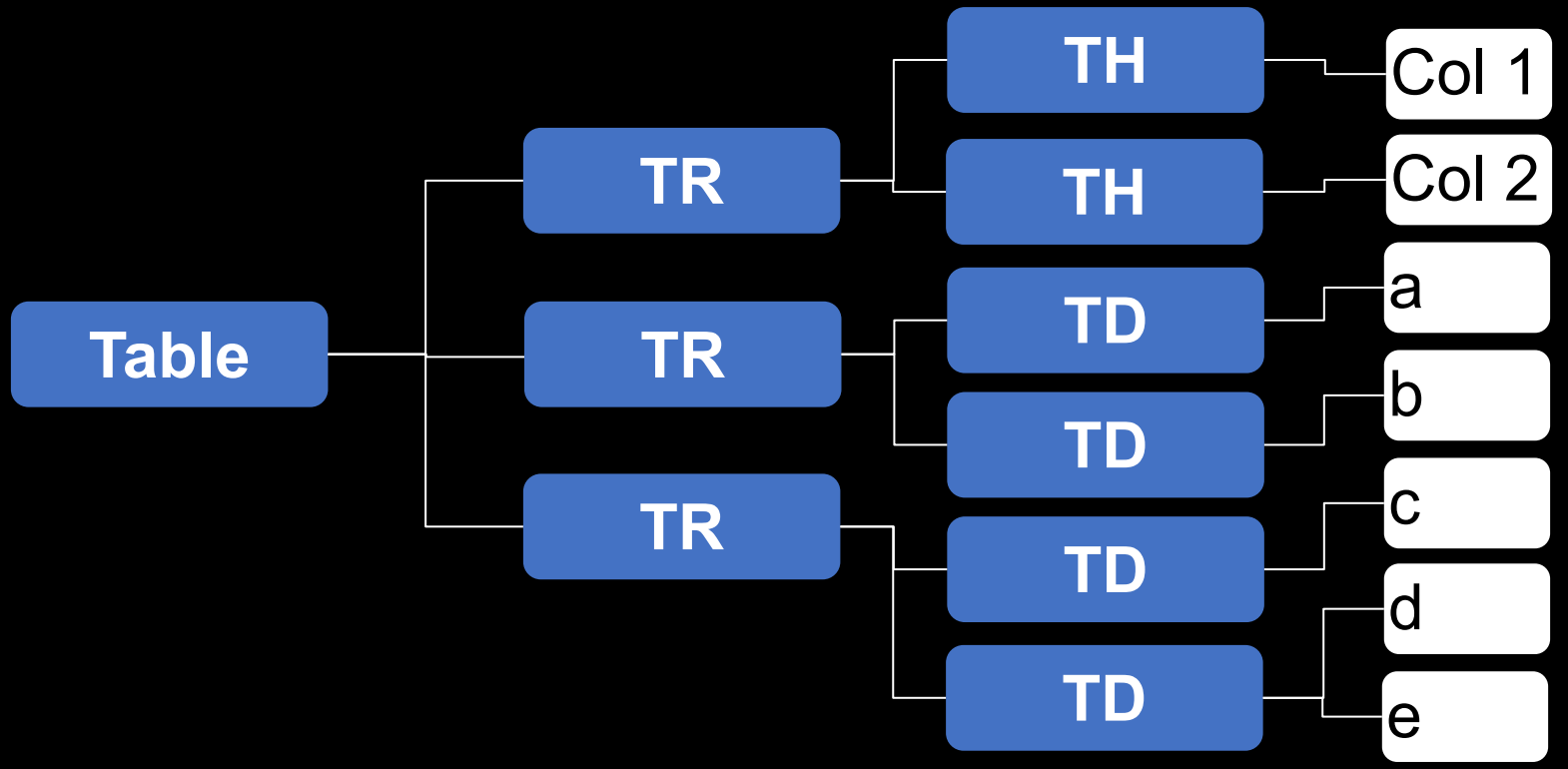
Lists – L tags

1. a
2. b
3. cd

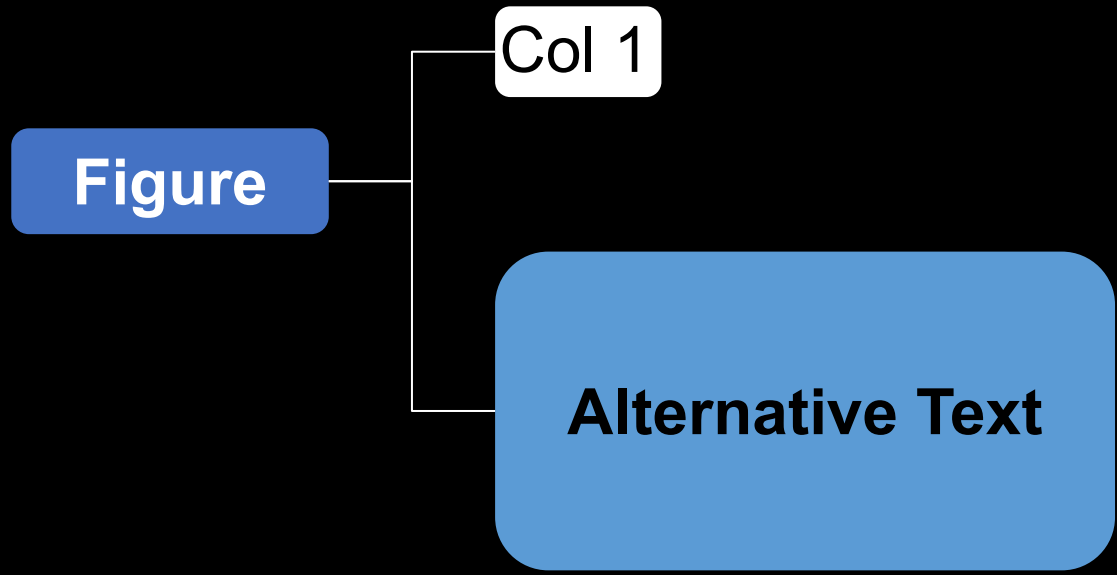


Tables – table tags

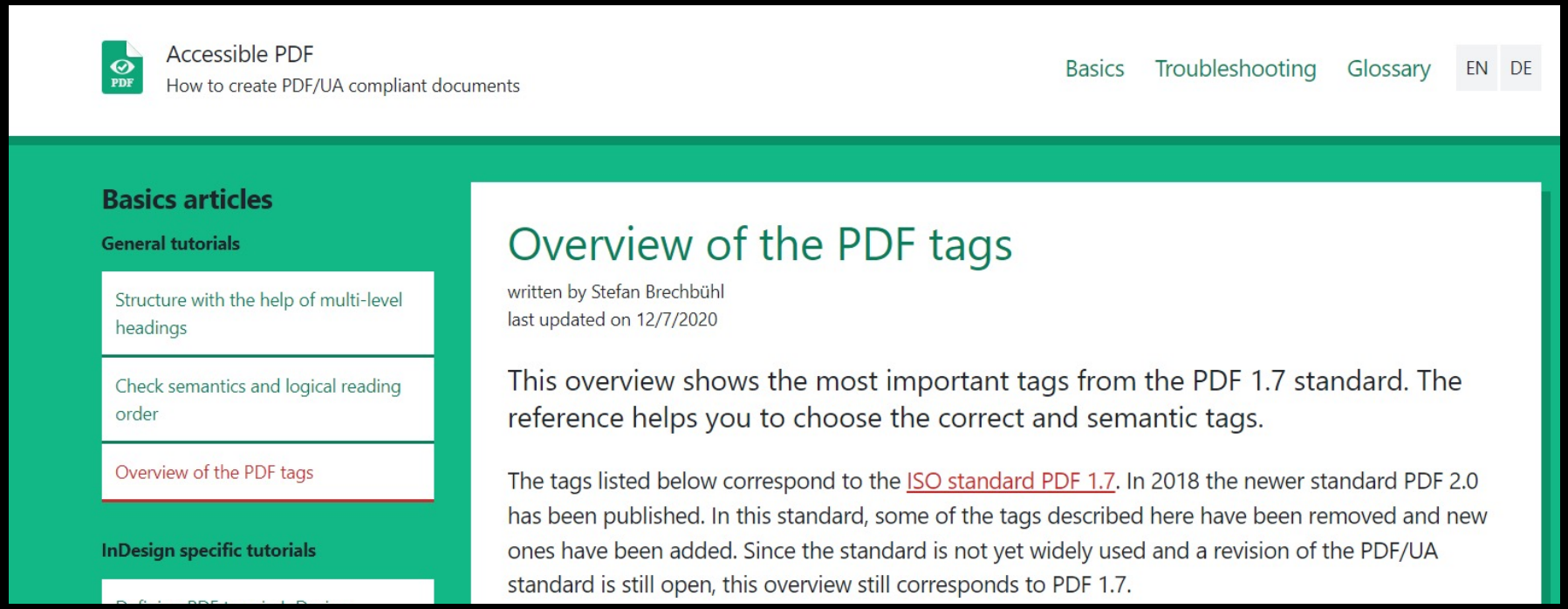
Col 1	Col 2
a	b
c	de



Figures – figure tags



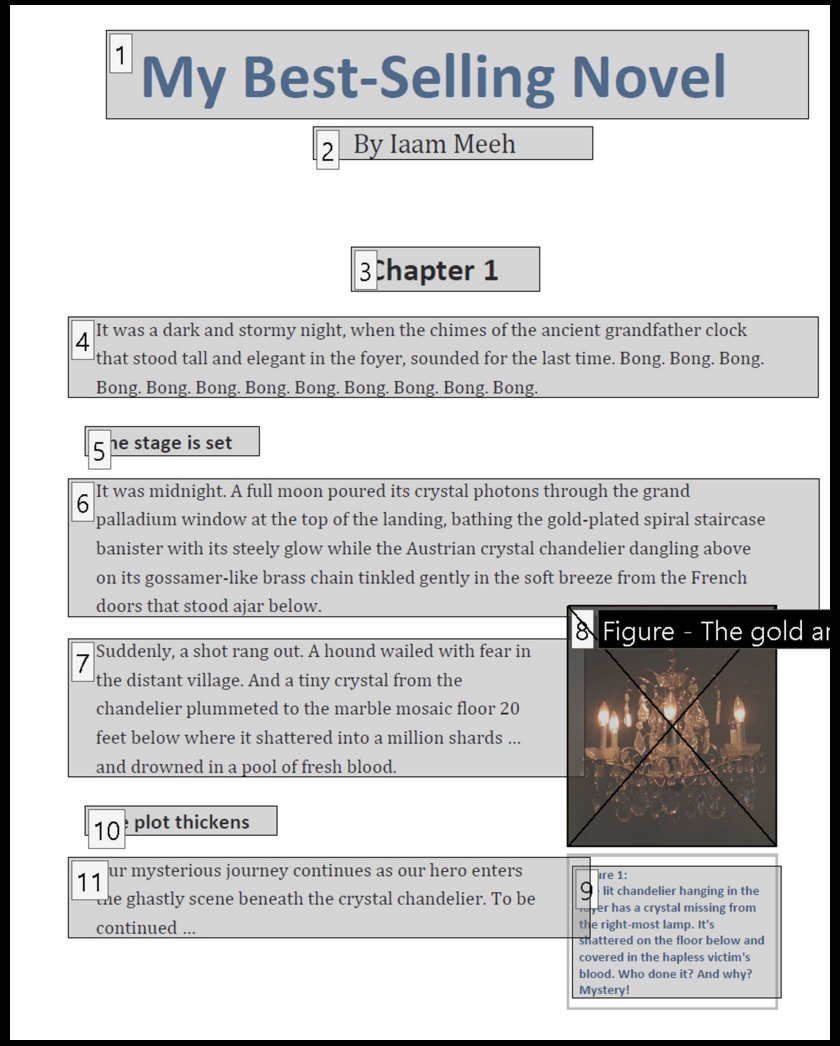
Other tags



Overview of the PDF Tags

→ https://accessible-pdf.info/

Reading Order



Source: https://www.pubcom.com/blog/2020_08-18_ReadingOrder/ordertree_01.png

Creating an accessible PDF

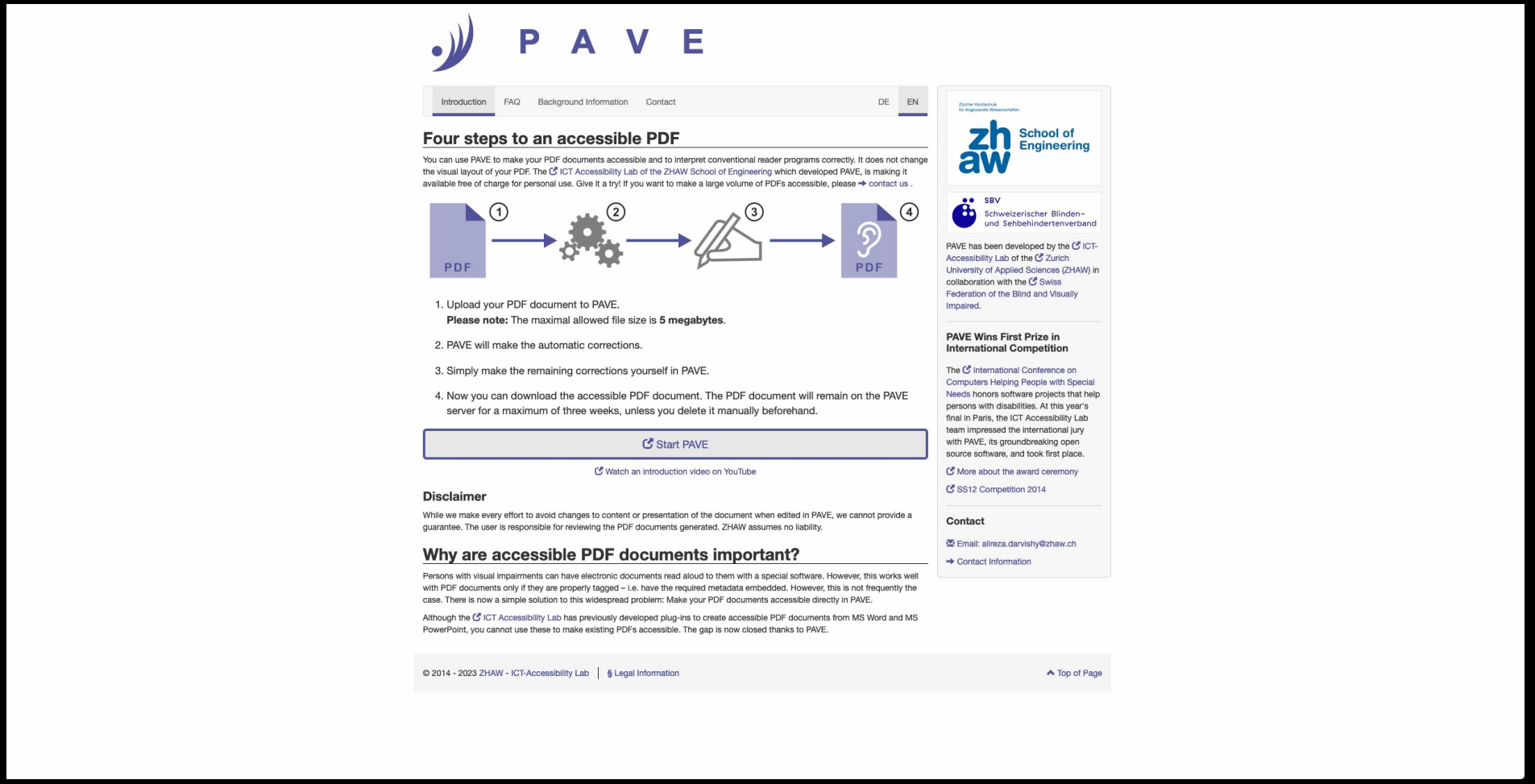
If the PDF exists:

- Adobe Acrobat Pro
- www.pave-pdf.org

If the source file exists:

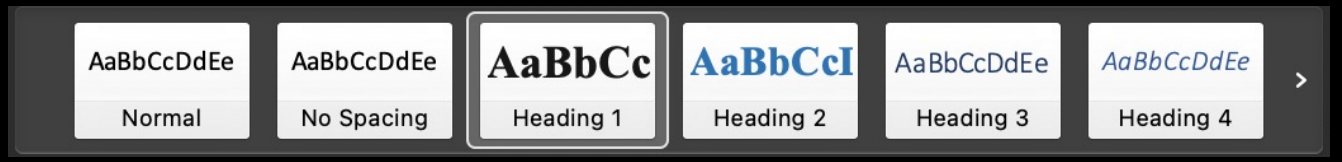
- Export the PDF with tags

PDF remediation with PAVE



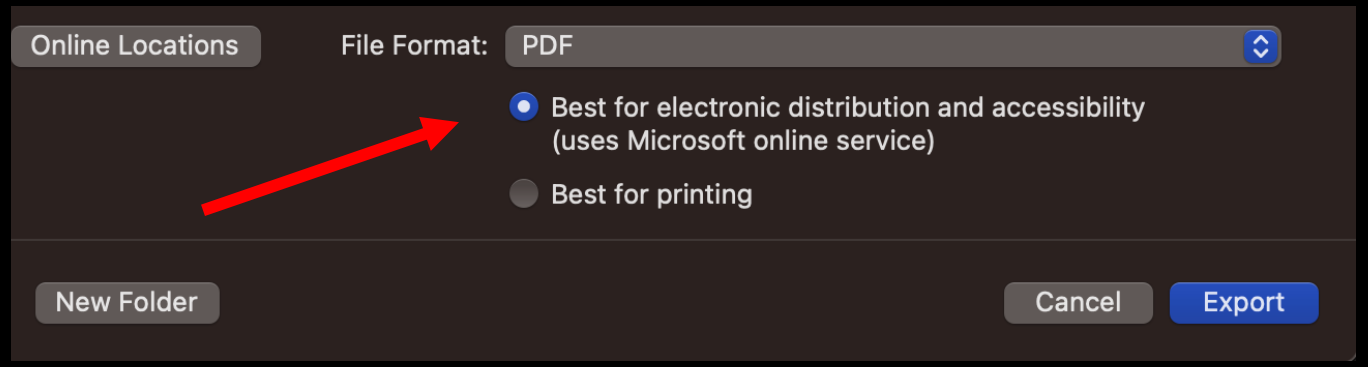
Export PDF with tags from Word

- Use the style templates



- Set alternative text

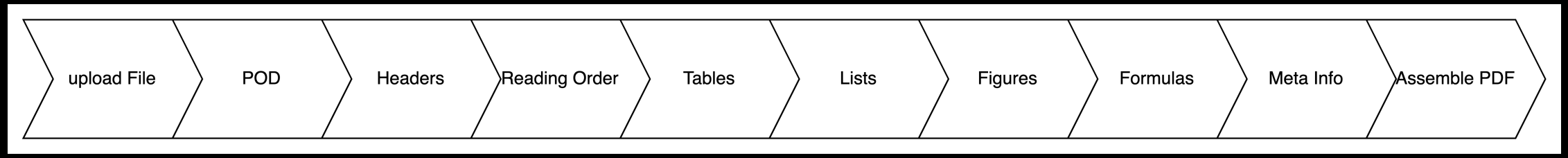
- Export the PDF with tags



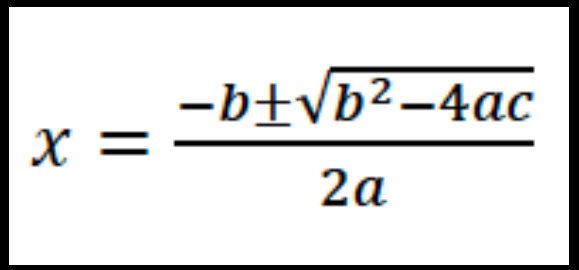
How AI can simplify PDF remediation

Project: Accessible Scientific PDFs for All

- AI allows automating remediation steps for complex documents



Alternative text for mathematical formulae



Alternative Text

x equals start fraction minus b plus-or-minus square root of b squared minus 4 a c end root over 2 a end fraction

Agenda

Part II: The state of scientific PDF accessibility in Switzerland

- Survey of accessibility in Swiss repositories
- What does the future look like?

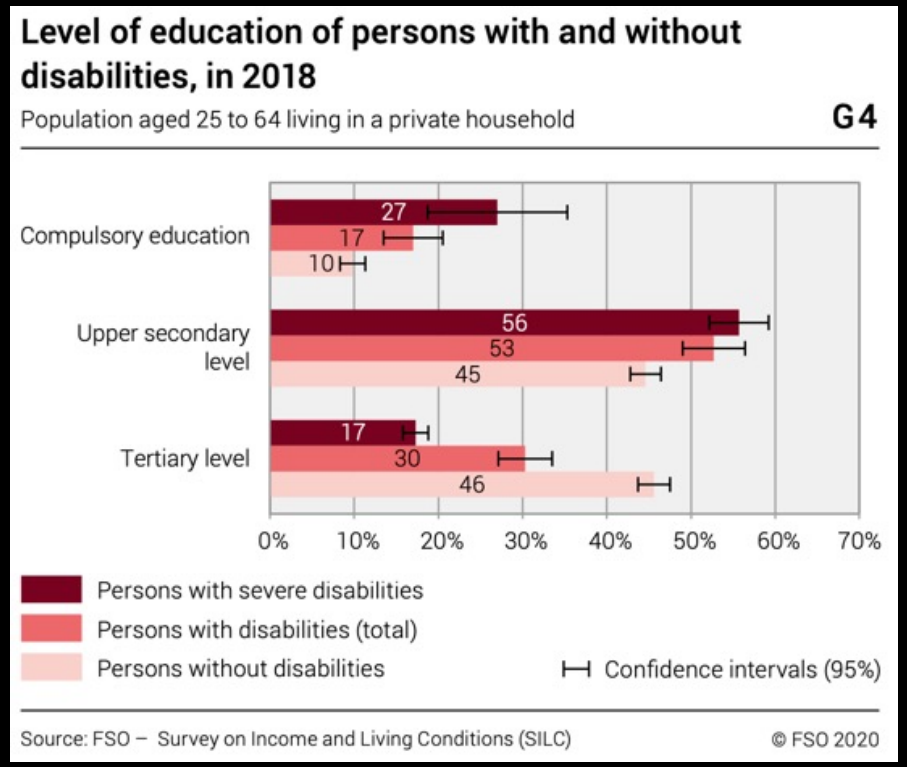
Motivation

People with disabilities in Switzerland

2018	As %	Number of persons
Persons with disabilities, severely limited	5.0%	347 000
Persons with disabilities, limited but not severely	17.2%	1 204 000
Total of persons with disabilities	22.2%	1 551 000

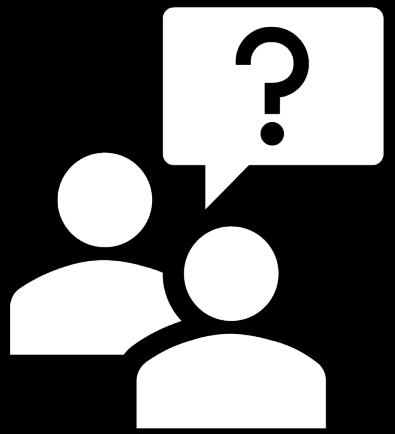
Source: FSO – Statistics on Income and Living Conditions (SILC) © FSO 2020

- Less people with disabilities pursue higher education than people without disabilities
- Around 30% of people with disabilities have a university level education



A mixed-method study on the state of scientific PDF accessibility in Swiss repositories

Quantitative analysis

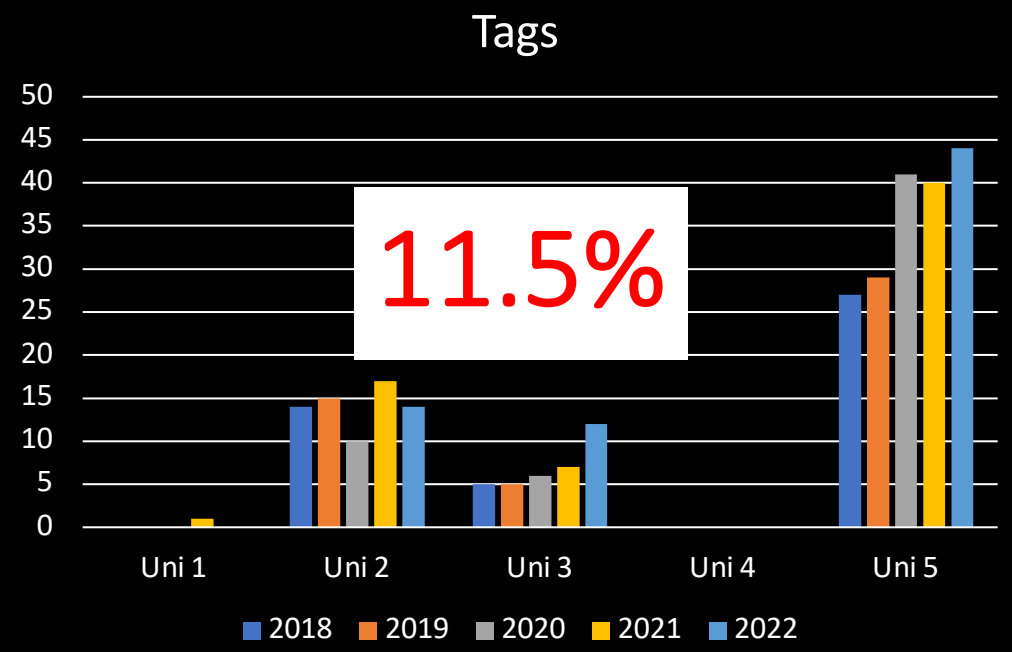


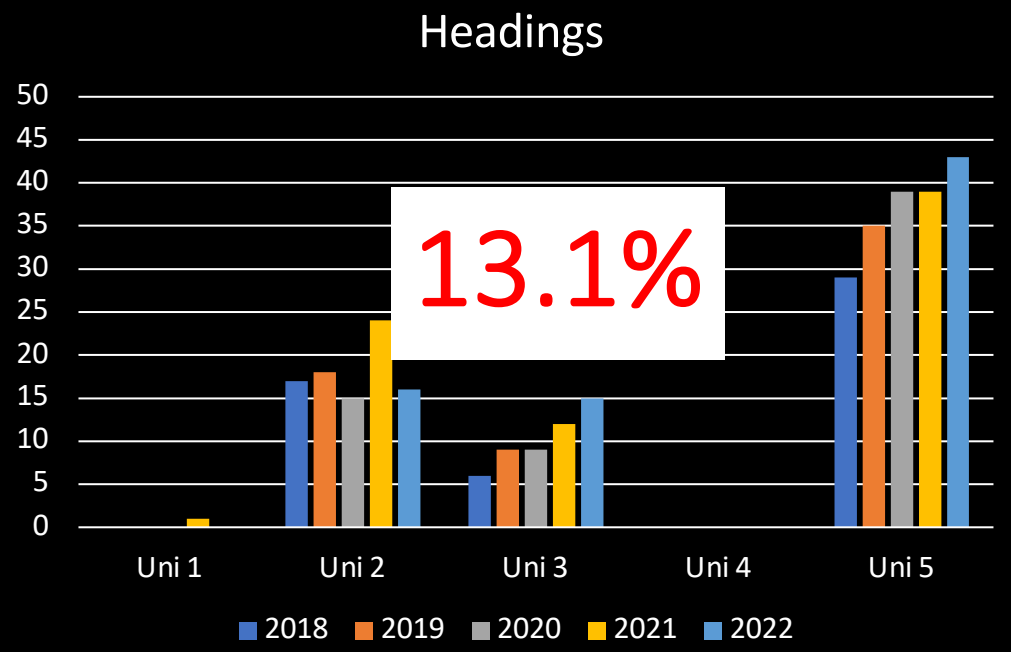
Qualitative analysis:
semi-structured interviews

Quantitative Analysis: Methods

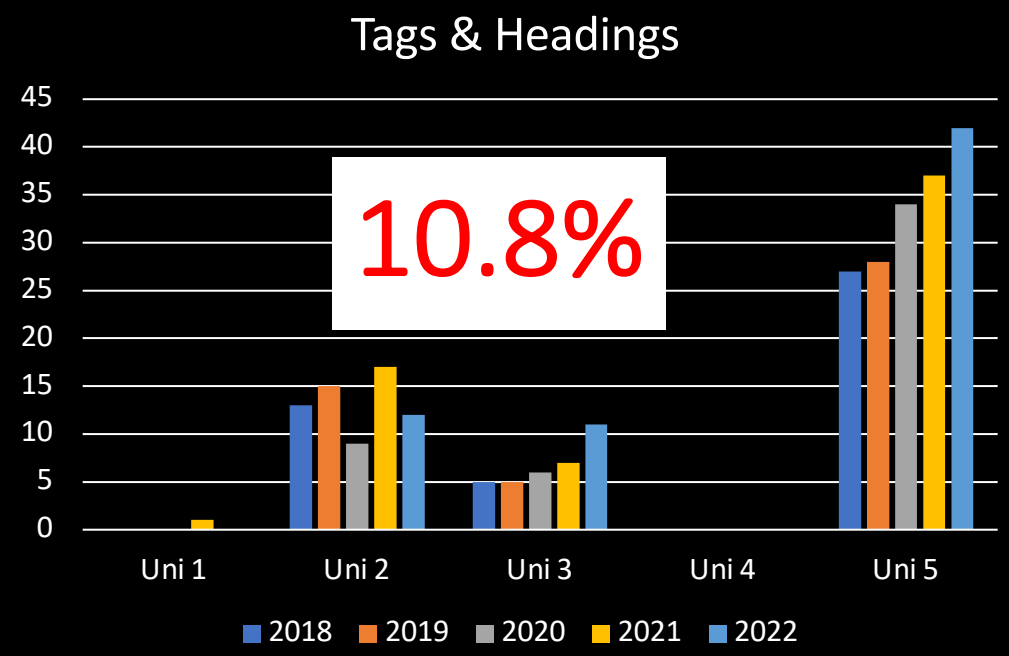
- 5 repositories of German-speaking Swiss universities
- Papers from 2018 – 2022
- All Open Access papers were downloaded
  - Random sample of 100 papers per year per university
    - = in total 2500 papers were analyzed
- 2 minimal accessibility features → tags & headings
- Automatically checked & counted

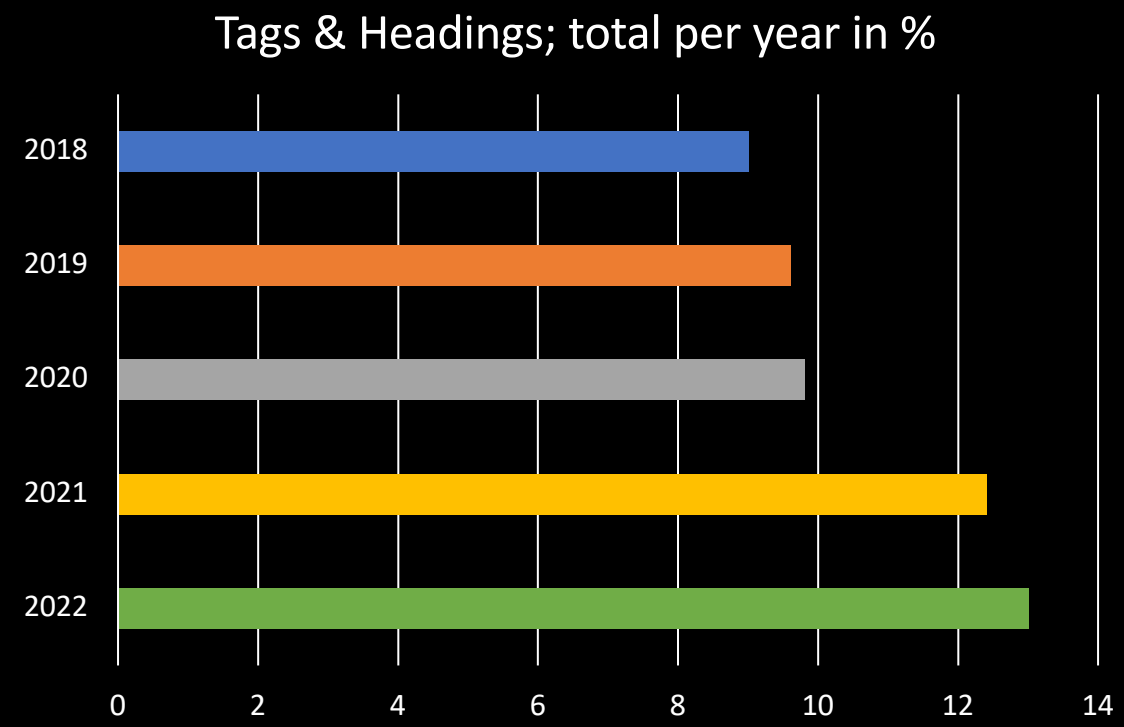
Quantitative Analysis: Results





Quantitative Analysis: Results





Quantitative Analysis: Discussion

- No university reaches a satisfactory % of accessible papers
- Big differences in % between universities
- Somewhat of an upward trend

- What could the repositories do additionally?
- Is the problem that there is a lack of knowledge or know-how?
- With what priority should this issue be addressed?

Qualitative Analysis: Methods

- 9 Swiss universities
- Semi-structured interviews with the heads of the libraries / repositories, managers of accessibility projects, etc.
- Anonymous participation
- Set of 12 questions in 4 topics
  - Knowledge & Opinions
  - Priority & Sensitization
  - Accessibility Measures
  - Future Plans

Qualitative Analysis: Methods

1) Could you please briefly introduce yourself and your role in the repository?

Knowledge and Opinions

2) What do you know about the topic of accessible PDF documents?
3) In your opinion, what is an accessible PDF document?
4) How would you rate the accessibility of the publications in your repository?
    a. Do you keep internal statistic on the accessibility of accessible publications?

Priority and Awareness

5) With what priority do you address the issue of accessibility in your repository?
6) How are the employees at the repository sensitized on the topic of "accessibility"?
    a. Are there trainings or conferences on the topic of "accessibility" for the employees?
    b. Is attendance in such trainings or conferences mandatory?
7) Is there any exchange between the repositories of universities in Switzerland – through associations, conferences, etc.)?

    If yes…

    a. To what extent is "accessibility" addressed in the discussions?

Accessibility Measures

8) What are specific measures underway in your repository to keep it as accessible as possible (in terms of accessibility of publications)?
    a. Do you have internal guidelines to ensure accessibility of published publications (mostly in PDF format)?
9) Do you check the accessibility of new incoming documents before publishing them on the repository?

    If yes…

    a. How do you check the accessibility of the documents (manually, etc.)?
    b. Since when do you check the accessibility of the documents?
10) How do you handle it when, if for example, a visually impaired person requests an accessible version of a document?
    a. Do you offer a service of providing alternative document formats?

Wishes and Conclusion

11) In your opinion, what fundamental change is needed at universities and research institutions for science in general to become more accessible?
12) What do you intend to do in the future to improve accessibility in your repository?

Qualitative Analysis: Results (1/2)

- Definition of ”accessible PDF document”: mainly focused on Open Access & structure
- Estimated amount of accessible documents = low
  - No statistic on the number of accessible documents
- Almost no repository provides trainings on accessibility for their employees
- Discussions within Open Access Working Group (Arbeitsgruppe)

  → Document accessibility is not a topic

Qualitative Analysis: Results (2/2)

- Accessibility has a low priority
  - Accessibility not seen as a responsibility
  - ”Resources” as a big hindrance
- No measures to improve document accessibility
  - Accessibility measures usually concern the platform and UI
  - Almost no repository has internal guidelines
    → One exception with dissertations
- No check of the accessibility of incoming documents
- Reactive approach → ready to remediate documents on a case-by-case basis, but no established services
- Future plans don’t include increasing accessibility of PDFs

Qualitative and quantitative Analysis: Combined Results

- Repositories of universities with accessibility guidelines had more PDFs with tags & headings
  - CAVE → concrete (!) guidelines necessary, i.e. all documents from within the university have to be accessible
- The more knowledge, the likelier it was for the repository to be willing to have more accessible PDFs
  - BUT willingness to be more accessible didn't translate into factual accessibility
- Certain common practices of some universities hinder creating accessible PDFs
  - e.g. using LaTeX

Recommendations

Seek out knowledge & raise awareness on accessibility of PDF documents
- →Include people with disabilities in decisions and proposals
- →Participate in trainings and conferences
- →Collaborate with other universities to establish accessibility goals

Make accessibility a requirement
- → Internal change: make accessibility a priority and a requirement (for authors, publishers, and repository managers)
- → Keep statistics on accessibility features of documents of own platform
- → Check accessibility of incoming documents before publication

Use and establish guidelines
- → Existing ones: ZHAW guidelines, Swissuniversities, Webaim, etc.
- → Build new guidelines within own institution

Don’t only rely on laws to push you to make a change! → Willingness and open mindset are important

Questions? Comments? Ideas?