Tag! Your PDF is It!

Alejandro Piñeiro and Joanmarie Diggs

GUADEC 2013

igalia

Free Software Engineering

Topics

- Tagged PDFs:
  - What They Are
  - Why We Want Them
  - How to Make Them
- Current Status of the Project
- Getting the Code (and what you'll see when you do)

Tagged PDFs

Tagged PDF > PDF

- Meta-information about page content
- HTMLish tags and IDs for text spans
- Alternative text for images
- Replacement text for symbols

Why We Want Them

- Enhanced document accessibility
- Through exposure of structural and semantic information associated with the tags

Thanks (again) Friends of GNOME!!!

Why We Want Them (cont.)

- Reflow functionality (e.g. for mobile devices)
- Export to other applications with format, layout, font data, etc.
- Copy and paste to other applications with some fundamental retention of content format

Making Tagged PDFs

- ✘ AbiWord: No
- ✘ Google Docs: No
- ✘ LaTeX: No
- ✘ Scribus: No
- ✘ PDF Studio: No
- ✘ python-pisa: No

- ✔ LibreOffice: Yes (and it's easy!)

PDF Options

General | Initial View | User Interface | Links | Security

Range

- All
- Pages
- Selection

Images

- Lossless compression
- JPEG compression
  - Quality: 90%
- Reduce image resolution: 300 DPI

Watermark

- Sign with Watermark

Watermark Text

General

- Embed OpenDocument file
  - Makes this PDF easily editable in LibreOffice
- PDF/A-1a
- Tagged PDF
- Create PDF form
  - Submit format: FDF
  - Allow duplicate field names
- Export bookmarks
- Export comments
- Export notes pages
- Export hidden pages
- Export automatically inserted blank pages
- Embed standard fonts
- View PDF after Export

Export | Cancel | Help

PDF/A-1a > Tagged PDF

- Objective: Search and repurpose document content

- Includes:
  - PDF/A-1b: Reproduce document appearance
  - Structure / Hierarchy
  - Tagged PDF
  - Unicode character maps
  - Language specification

Current Status

Tagged PDF Support

- ✔ Parse the document structure tree: Poppler
- ✔ Expose the tree and attributes: Poppler GLib
- ✔ Provide tools to examine and verify result: Poppler
- Create parallel object tree with attributes: Evince
- (Expose object tree and attributes via ATK: Evince)

PDF/A-1a Support

- ? PDF/A-1b
- ✔ Tagged PDF
- ✔ Structure / Hierarchy
- ? Unicode character maps
- ✔ Language specification

What's Next?

- Create parallel object tree with attributes: Evince
- (Expose object tree and attributes via ATK: Evince)

? PDF/A-1b and Unicode character maps

? Adding support to LaTeX, et al.

Getting the Code
(and what you'll see when you do)

Credit Where Credit is Due

- Adrián Pérez: Document Parser Extraordinaire
- Carlos García Campos: Maintains Evince & Poppler

Thanks Guys!!!

Getting the Code

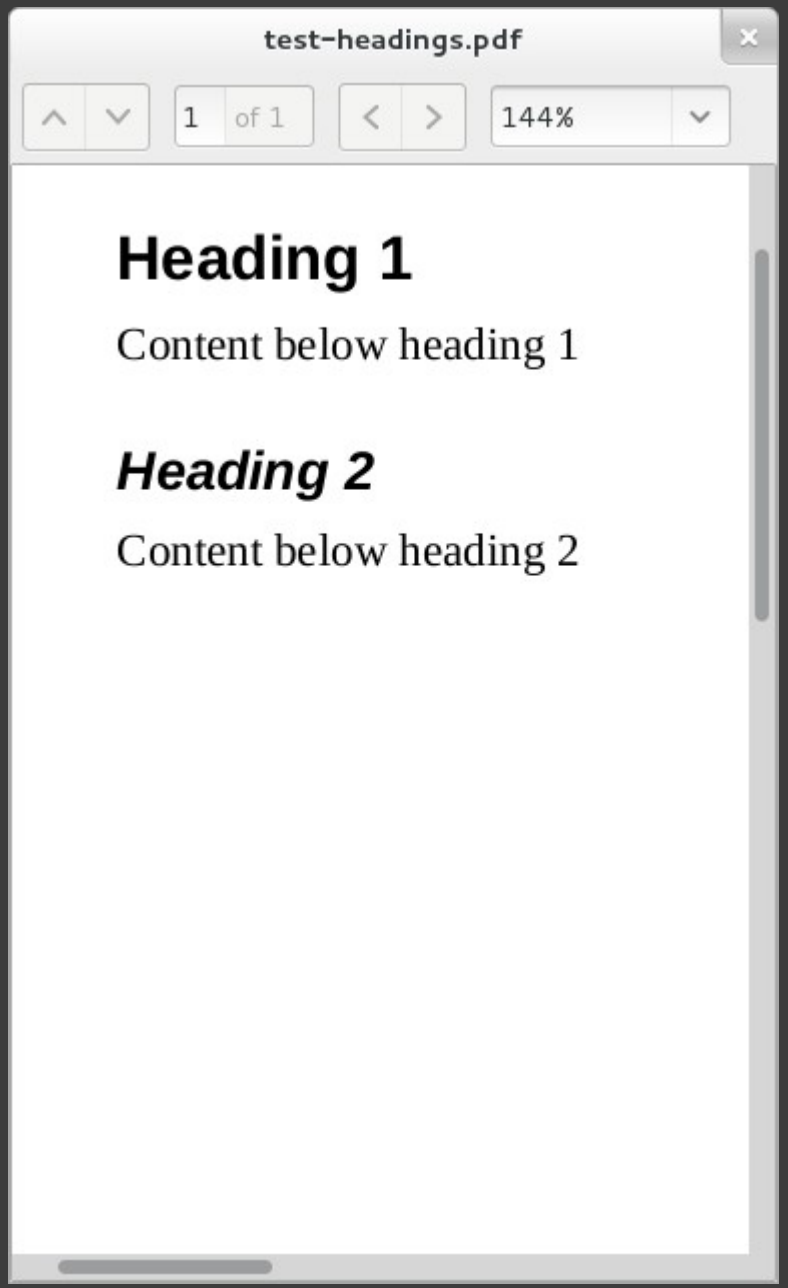
- git://git.freedesktop.org/git/poppler/poppler
- *Today*
  - Branch: tagged-pdf
  - Patches: fdo bugs 64816 and 67710
- *Soon:* master branch

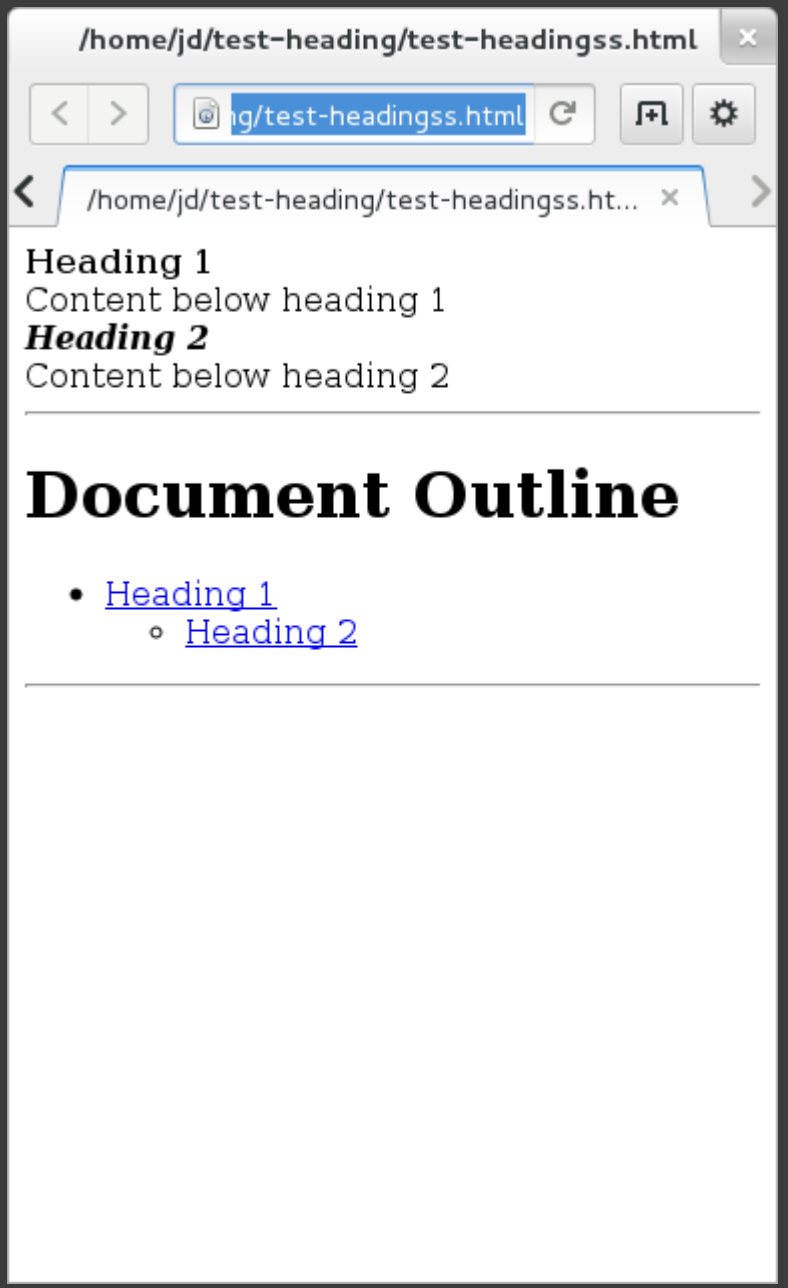
Getting the Code (cont.)

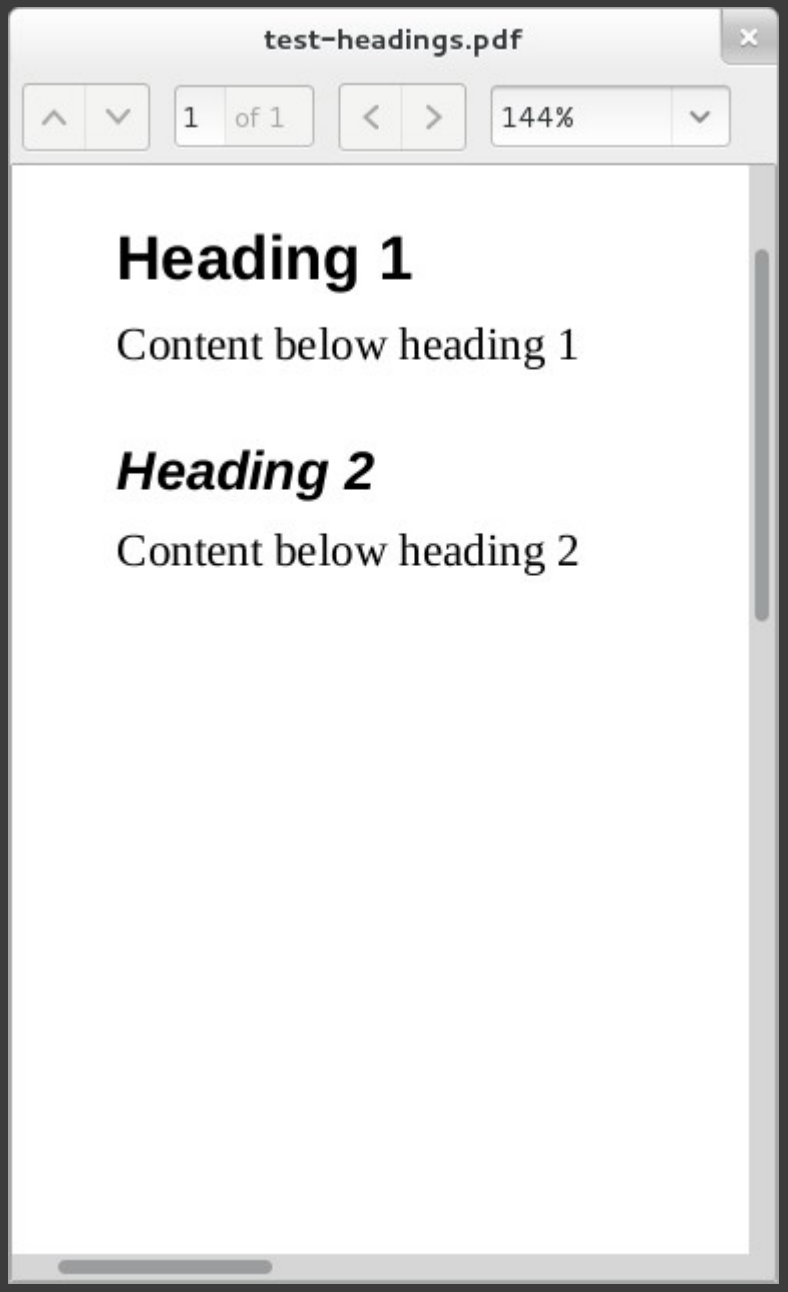
- Poppler:
  10 files changed, 2309 insertions(+), 17 deletions(-)
- Popper Glib:
  16 files changed, 3011 insertions(+)
- Utils:
  3 files changed, 661 insertions(+), 2 deletions(-)

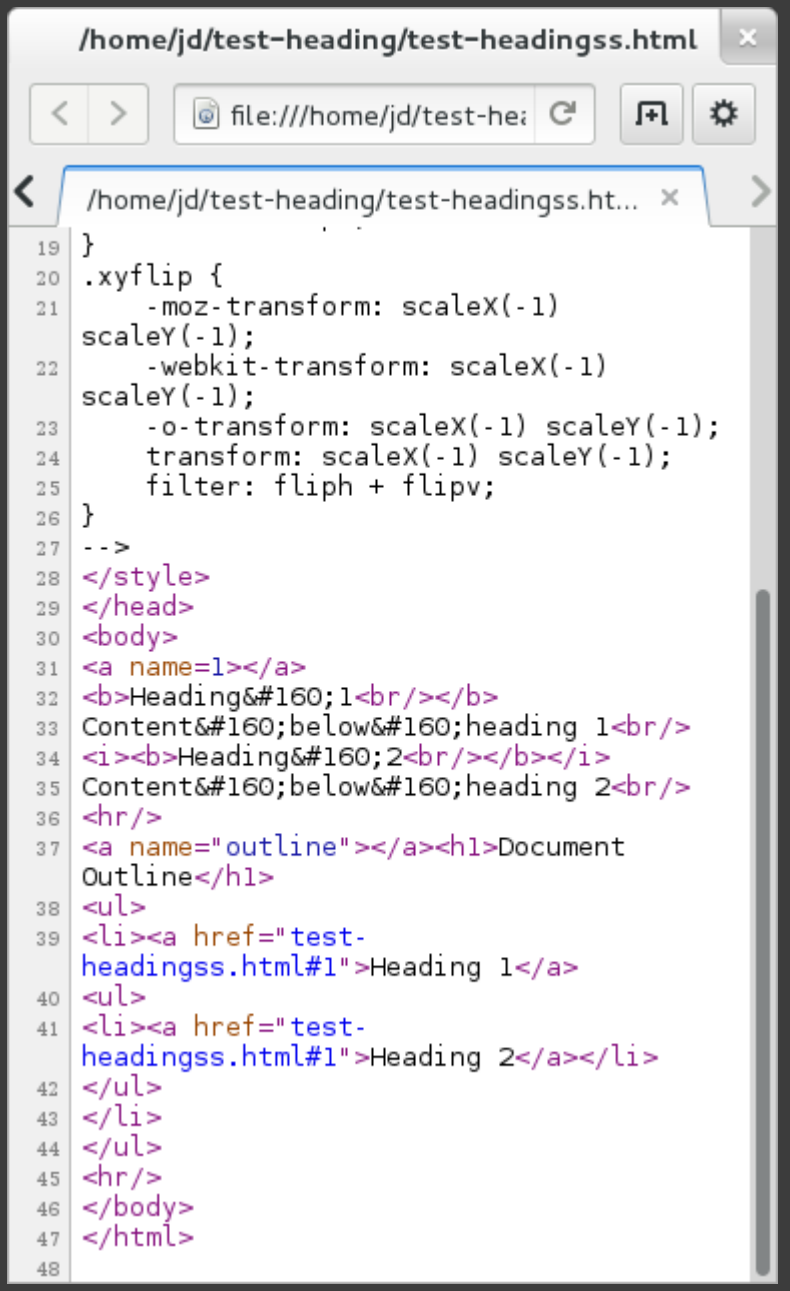
Associated Output Tools: Before

- pdfinfo: author, editor, etc.
- pdftotext: content (plain text)
- pdftohtml: content (barely formatted text)


test-headings.pdf
1 of 1
144%
Heading 1
Content below heading 1
Heading 2
Content below heading 2


/home/jd/test-heading/test-headingss.html
g/test-headingss.html
/home/jd/test-heading/test-headingss.ht...
Heading 1
Content below heading 1
Heading 2
Content below heading 2
Document Outline
Heading 1
Heading 2


test-headings.pdf
1 of 1
144%
Heading 1
Content below heading 1
Heading 2
Content below heading 2


/home/jd/test-heading/test-headingss.html
file:///home/jd/test-hea
/home/jd/test-heading/test-headingss.ht...
}
.xyflip {
-moz-transform: scaleX(-1) scaleY(-1);
-webkit-transform: scaleX(-1) scaleY(-1);
-o-transform: scaleX(-1) scaleY(-1);
transform: scaleX(-1) scaleY(-1);
filter: fliph + flipv;
}
-->
</style>
</head>
<body>

Heading 1
Content below heading 1
Heading 2
Content below heading 2
<hr/>
<h1>Document Outline</h1>

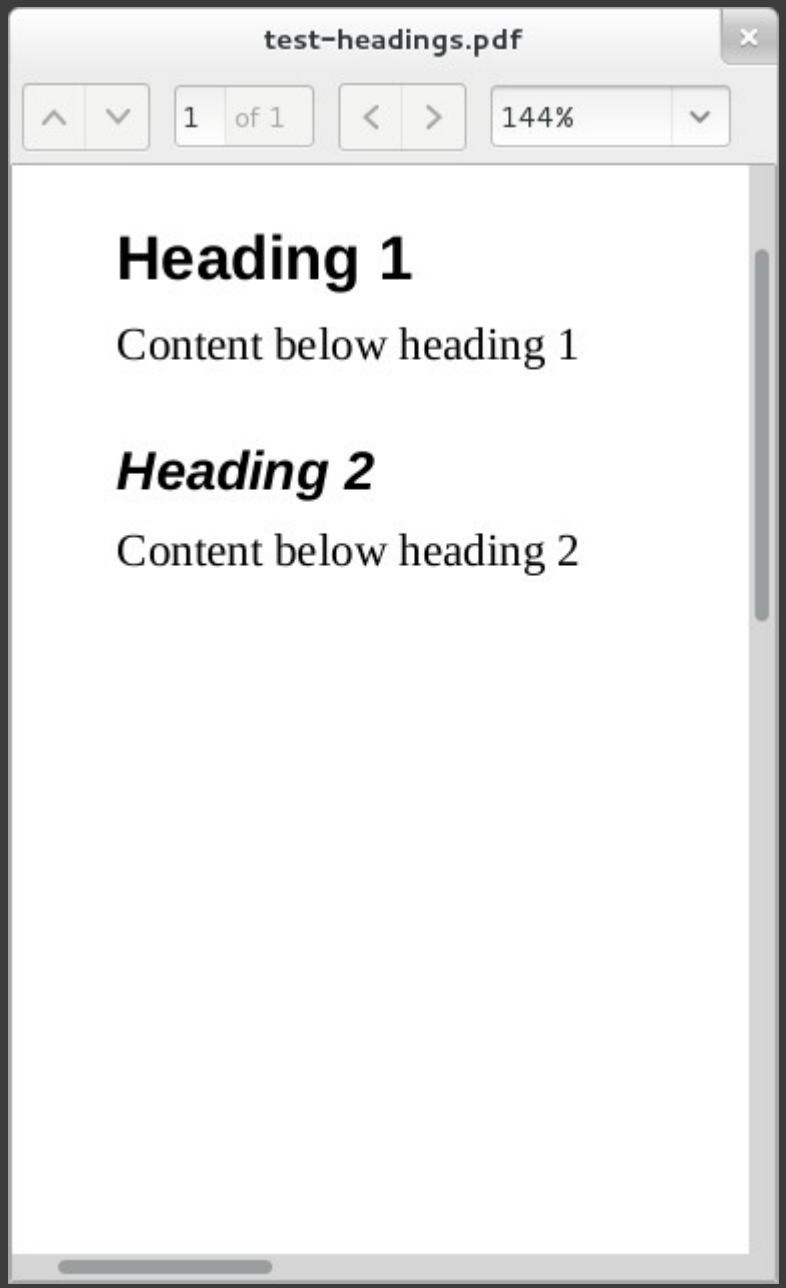
Heading 1

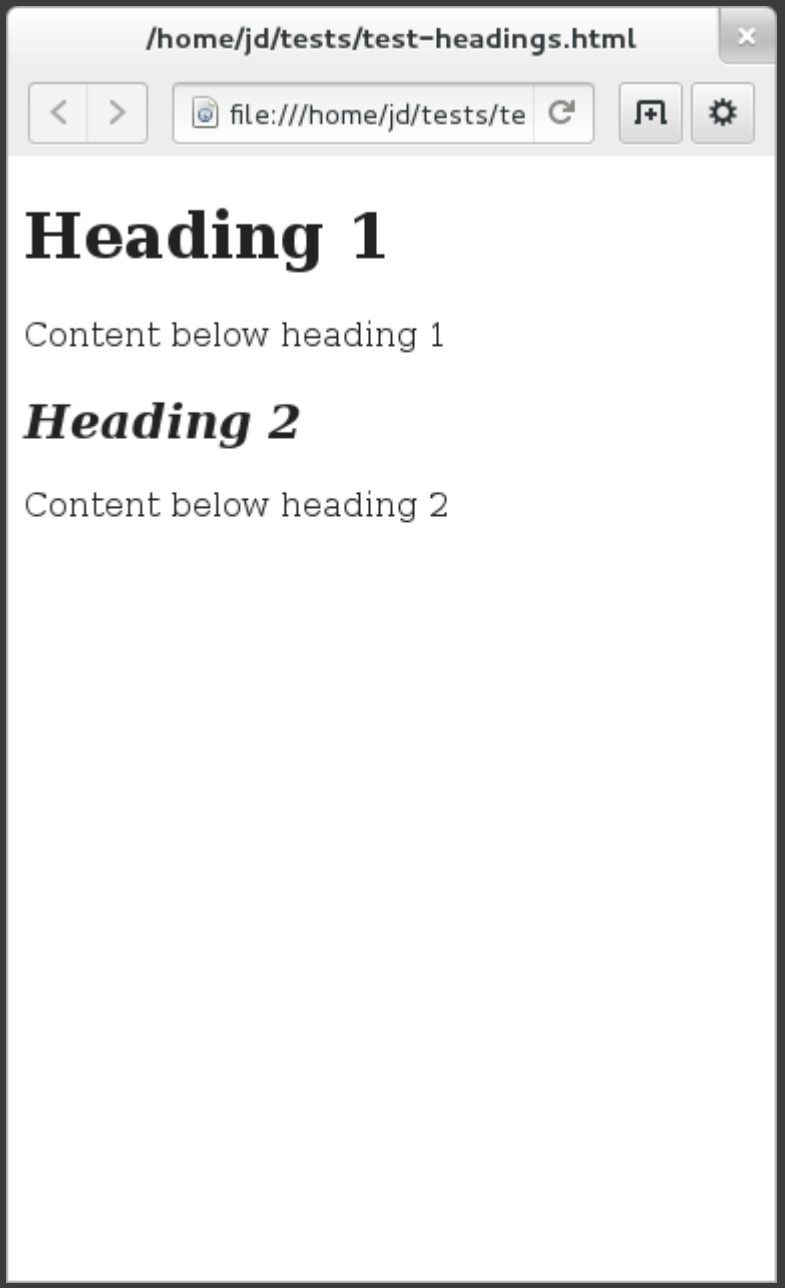
Heading 2

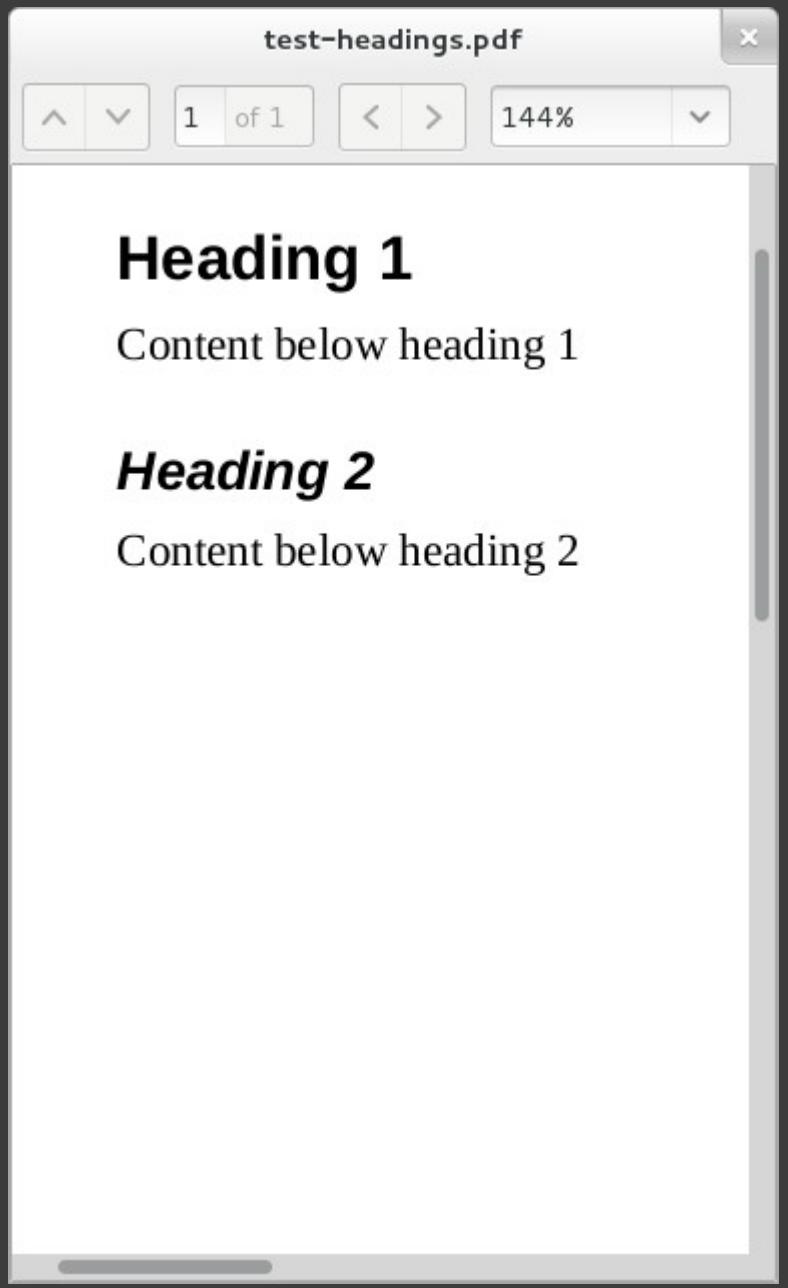
<hr/>
</body>
</html>

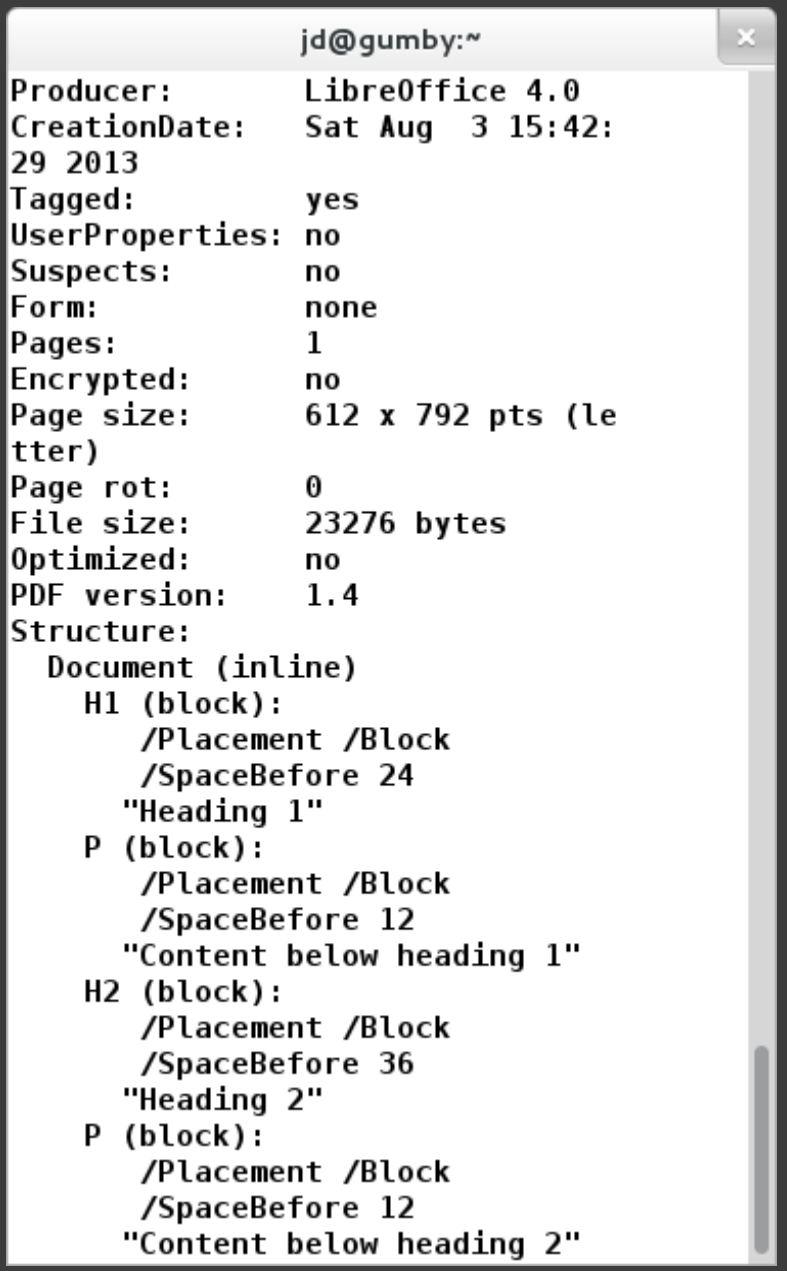
Associated Output Tools: After

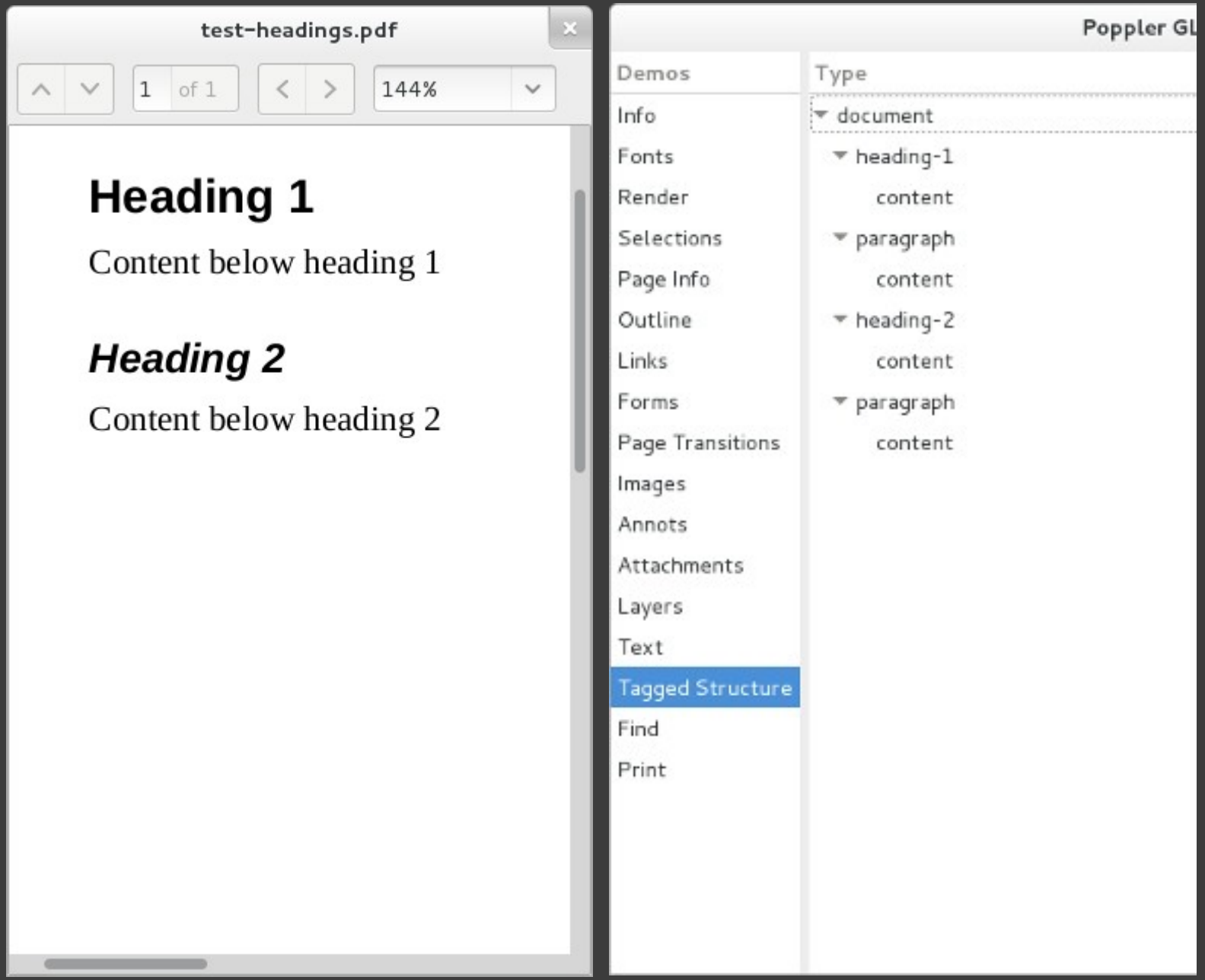
- pdfstructtohtml: like pdftohtml but preserves tags
- pdfinfo's new options:
  - hierarchy
  - hierarchy along with content of each element
- poppler-glib-demo: new option to display hierarchy

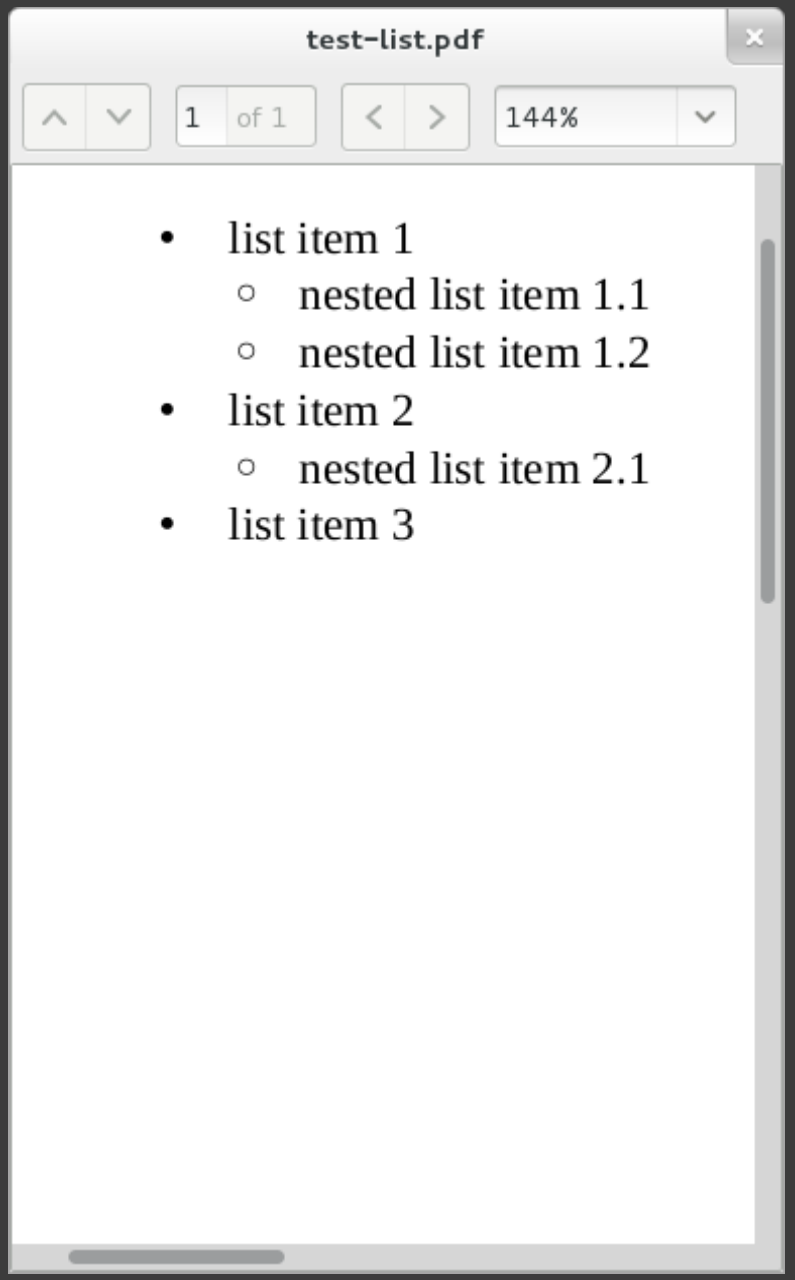

test-headings.pdf
1 of 1
144%
Heading 1
Content below heading 1
Heading 2
Content below heading 2

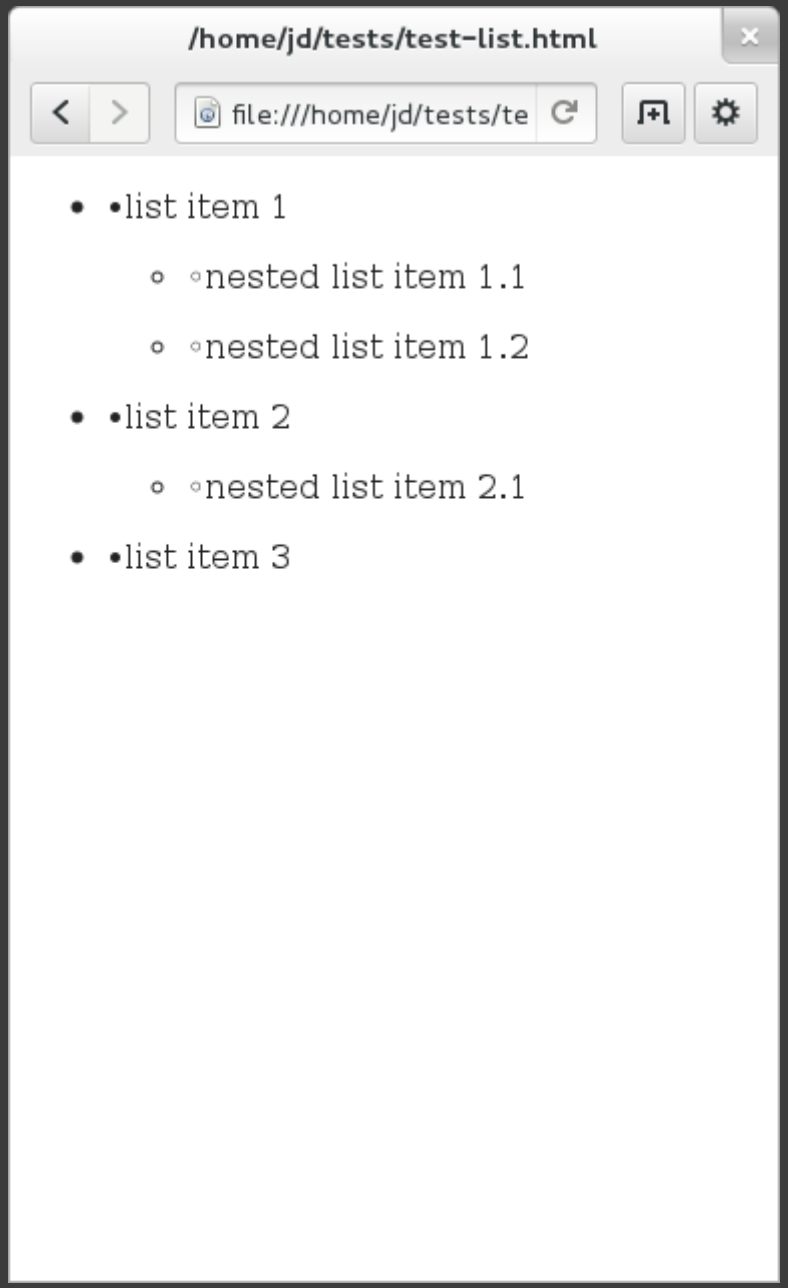

/home/jd/tests/test-headings.html
file:///home/jd/tests/te
Heading 1
Content below heading 1
Heading 2
Content below heading 2

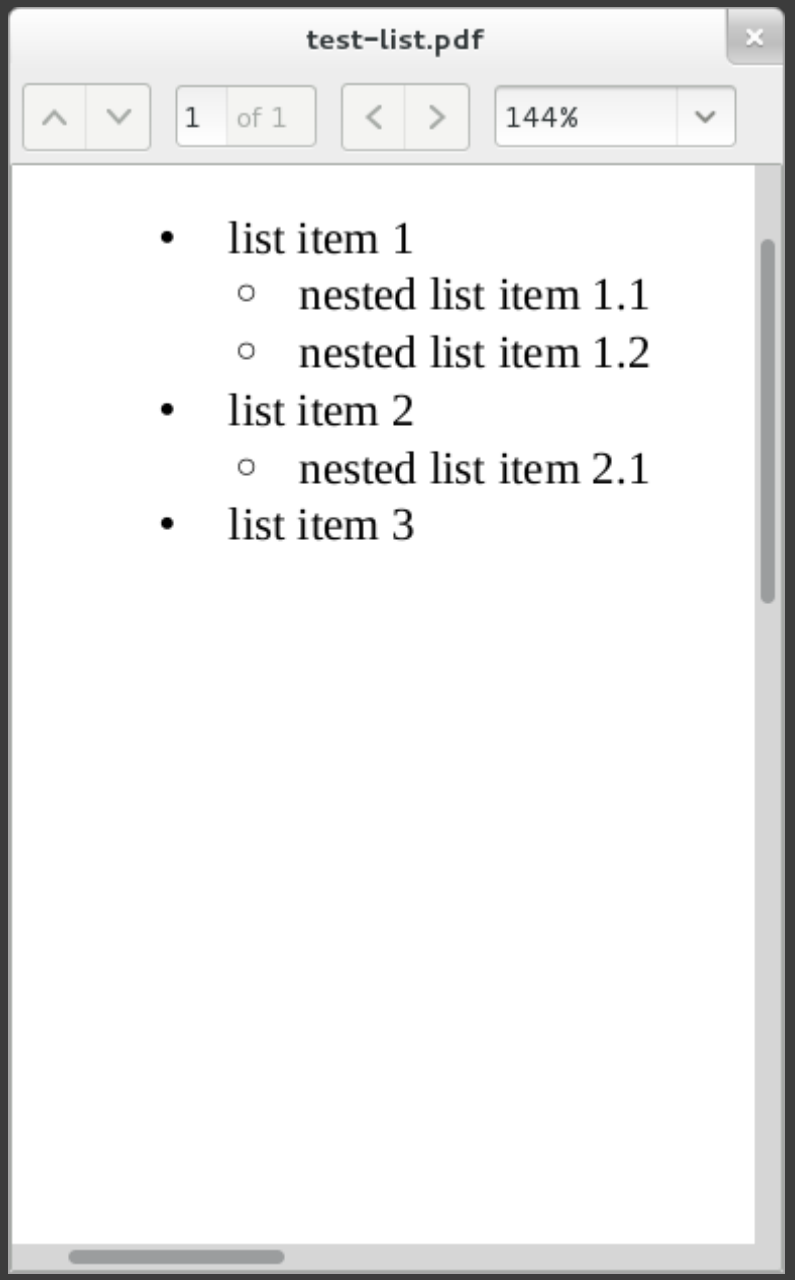

test-headings.pdf
1 of 1
144%
Heading 1
Content below heading 1
Heading 2
Content below heading 2

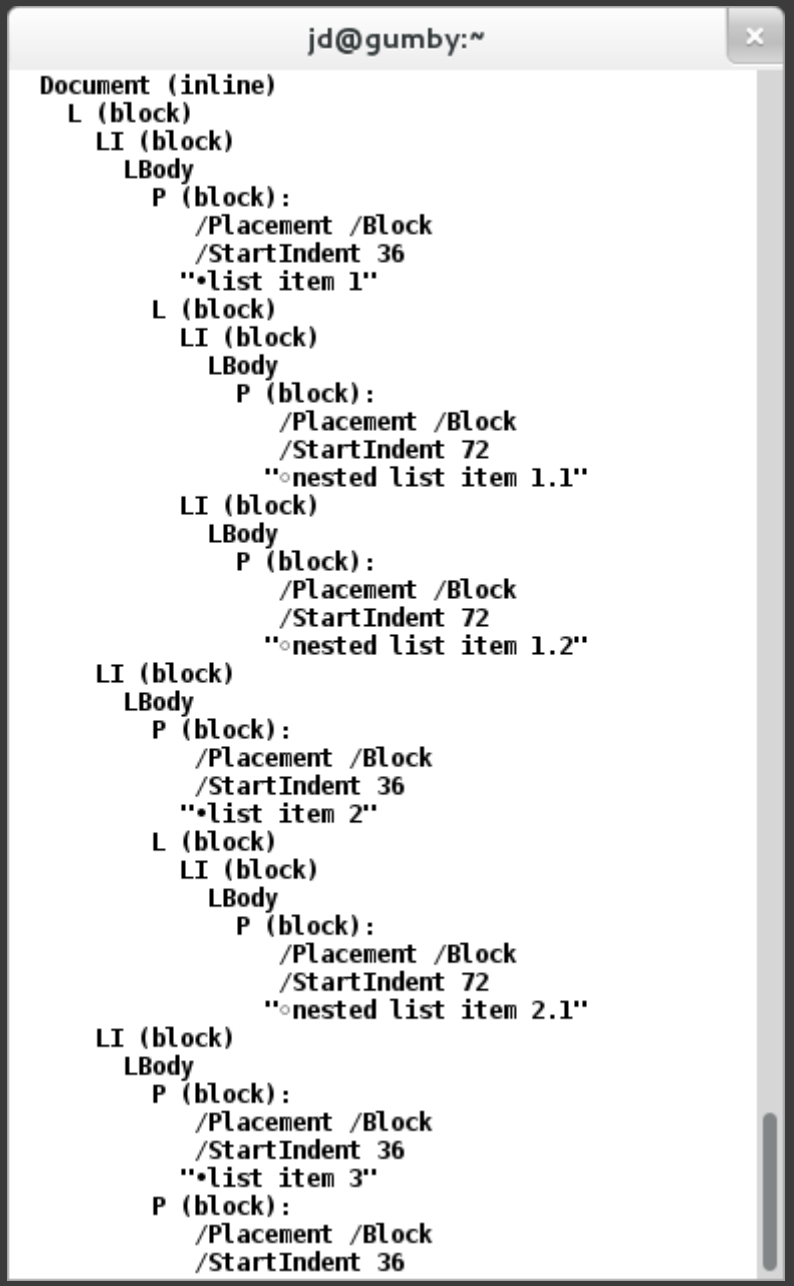

jd@gumby:~
Producer: LibreOffice 4.0
CreationDate: Sat Aug 3 15:42:
29 2013
Tagged: yes
UserProperties: no
Suspects: no
Form: none
Pages: 1
Encrypted: no
Page size: 612 x 792 pts (le
tter)
Page rot: 0
File size: 23276 bytes
Optimized: no
PDF version: 1.4
Structure:
Document (inline)
H1 (block):
/Placement /Block
/SpaceBefore 24
"Heading 1"
P (block):
/Placement /Block
/SpaceBefore 12
"Content below heading 1"
H2 (block):
/Placement /Block
/SpaceBefore 36
"Heading 2"
P (block):
/Placement /Block
/SpaceBefore 12
"Content below heading 2"

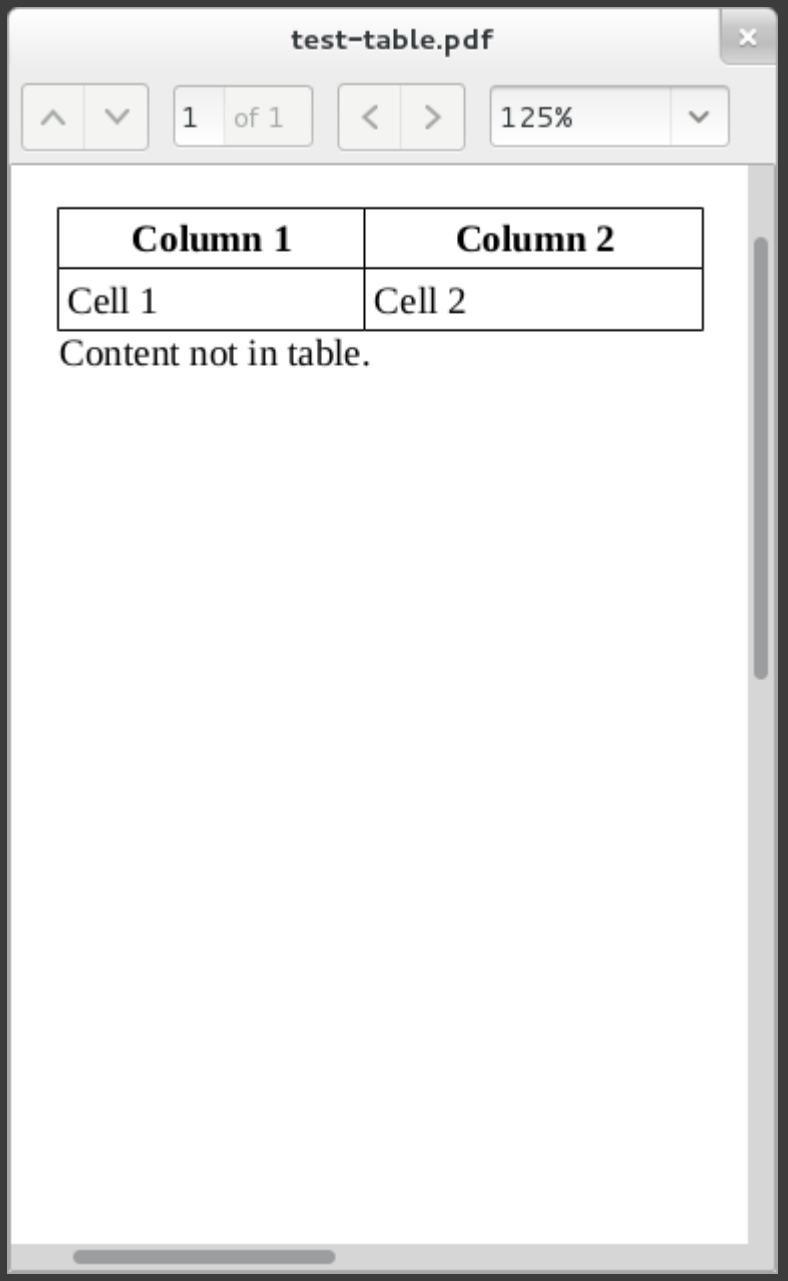

test-headings.pdf
1 of 1
144%
Heading 1
Content below heading 1
Heading 2
Content below heading 2
Poppler GL
Demos
Info
Fonts
Render
Selections
Page Info
Outline
Links
Forms
Page Transitions
Images
Annots
Attachments
Layers
Text
Tagged Structure
Find
Print
Type
document
heading-1
content
paragraph
content
heading-2
content
paragraph
content

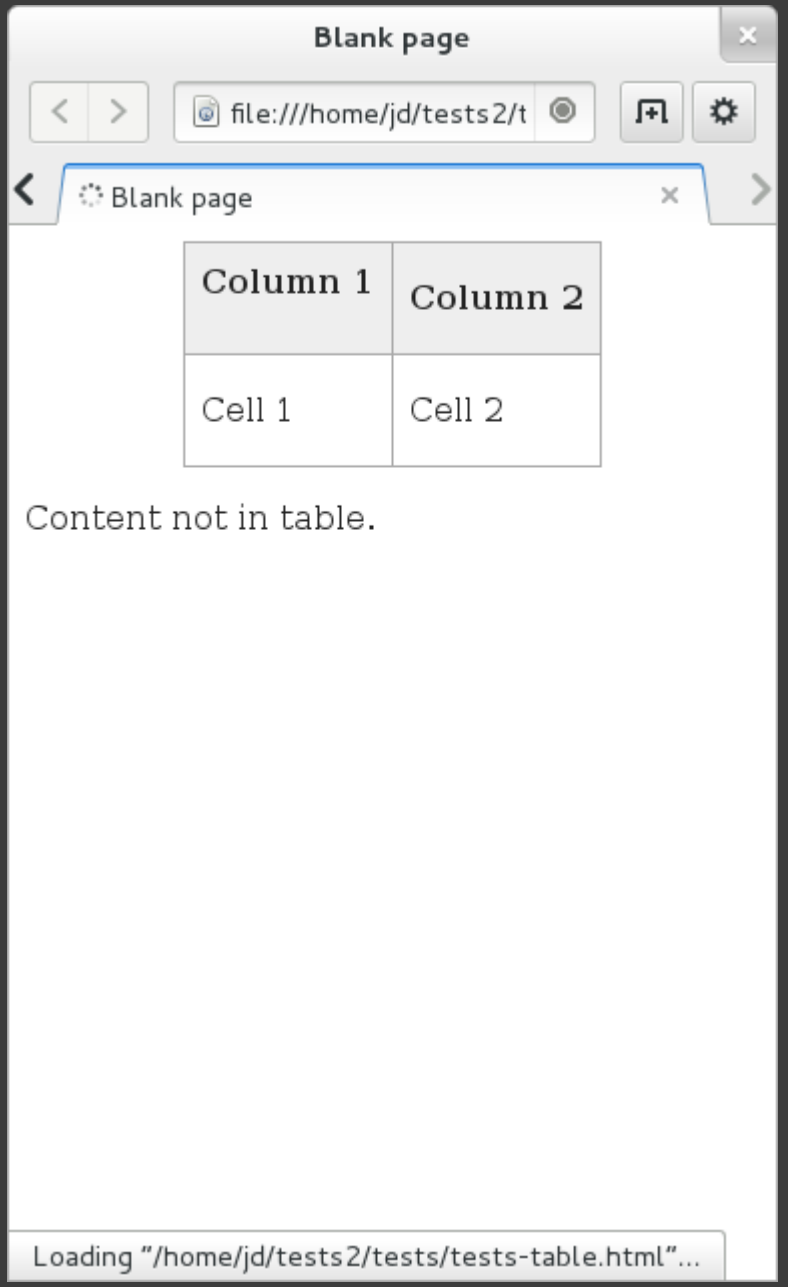

test-list.pdf
1 of 1
144%
list item 1
nested list item 1.1
nested list item 1.2
list item 2
nested list item 2.1
list item 3


/home/jd/tests/test-list.html
file:///home/jd/tests/te
•list item 1
◦nested list item 1.1
◦nested list item 1.2
•list item 2
◦nested list item 2.1
•list item 3


test-list.pdf
1 of 1
144%
list item 1
nested list item 1.1
nested list item 1.2
list item 2
nested list item 2.1
list item 3


jd@gumby:~
Document (inline)
L (block)
LI (block)
LBody
P (block):
/Placement /Block
/StartIndent 36
"•list item 1"
L (block)
LI (block)
LBody
P (block):
/Placement /Block
/StartIndent 72
"◦nested list item 1.1"
LI (block)
LBody
P (block):
/Placement /Block
/StartIndent 72
"◦nested list item 1.2"
LI (block)
LBody
P (block):
/Placement /Block
/StartIndent 36
"•list item 2"
L (block)
LI (block)
LBody
P (block):
/Placement /Block
/StartIndent 72
"◦nested list item 2.1"
LI (block)
LBody
P (block):
/Placement /Block
/StartIndent 36
"•list item 3"
P (block):
/Placement /Block
/StartIndent 36


test-table.pdf
1 of 1
125%
Column 1
Column 2
Cell 1
Cell 2
Content not in table.


Blank page
file:///home/jd/tests2/t
Blank page
Column 1
Column 2
Cell 1
Cell 2
Content not in table.
Loading "/home/jd/tests2/tests/tests-table.html"...

test-table.pdf

1 of 1 125%

Column 1	Column 2
Cell 1	Cell 2

Content not in table.

jd@gumby:~

```
Document (inline)
  Table (block):
     /Placement /Block
     /SpaceAfter 0.1
     /StartIndent -0.9
     /EndIndent 591.6
     /Width 997.2
     /Height 77.5
     /BBox [56.7 696.5 555.3 735.3]
    TR:
       /Placement /Block
      TH:
         /Placement /Inline
         /Width 193.5
         /Height 38.7
        P (block):
           /Placement /Block
           /TextAlign /Center
          "Column 1"
      TH:
         /Placement /Inline
         /Width 213
         /Height 38.7
        P (block):
           /Placement /Block
           /TextAlign /Center
          "Column 2"
    TR:
       /Placement /Block
      TD:
         /Placement /Inline
         /Width 193.5
         /Height 38.7
        P (block):
           /Placement /Block
          "Cell 1"
      TD:
         /Placement /Inline
         /Width 213
         /Height 38.7
        P (block):
           /Placement /Block
          "Cell 2"
```

Questions?