

DocEng 2018

2018-08-30, Halifax, NS, Canada

diffi: diff improved a preview

Gioele Barabucci

gioele.barabucci@uni-koeln.de

University of Cologne, Cologne Center for eHumanities

Universität zu Köln



```
$ echo "Hello world!" > file.txt  
$ gzip -c file.txt > file.txt.gz
```

```
$ echo "Hello world!" > file.txt  
$ gzip -c file.txt > file.txt.gz
```

```
$ diff file.txt file.txt.gz  
Binary files file.txt and  
file.txt.gz differ
```

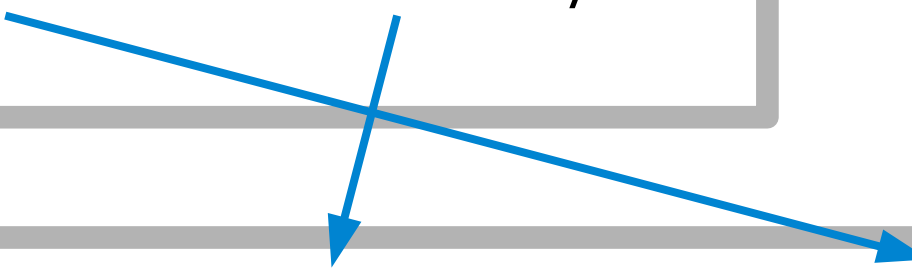
```
$ echo "Felicità!" > file.utf-8  
$ iconv -f utf-8 -t utf-16 \  
    < file.utf-8 > file.utf-16
```

```
$ echo "Felicità!" > file.utf-8  
$ iconv -f utf-8 -t utf-16 \  
    < file.utf-8 > file.utf-16
```

```
$ diff file.utf-8 file.utf-16  
Binary files file.utf-8 and  
file.utf-16 differ
```

```
<record color="red" value="1" />
```

```
<record color="red" value="1" />
```

Two blue arrows originate from the first code snippet. One arrow points from the 'color' attribute to the 'value' attribute in the second snippet. The other arrow points from the 'value' attribute in the first snippet to the 'color' attribute in the second snippet, illustrating a swap in attribute order.

```
<record value="1" color="red" />
```

```
<record color="red" value="1" />
```

The diagram illustrates a transformation of an XML element. A box at the top contains the original XML: `<record color="red" value="1" />`. Two blue arrows originate from this box: one points to the `value="1"` attribute in a box below, and the other points to the `color="red"` attribute in the same box. This indicates an attribute swap. Below the transformed XML box, two black arrows point down to a terminal window. The left arrow originates from the original box, and the right arrow originates from the transformed box. The terminal window shows the command `$ cmp r1.xml r2.xml` and the output `r1.xml r2.xml differ: byte 9, line 1`, where the file names are red.

```
<record value="1" color="red" />
```

```
$ cmp r1.xml r2.xml  
r1.xml r2.xml differ: byte 9, line 1
```

diffi: diff improved

A multi-format comparison tool that compares content.

CMV+P

What is a document?

Levels of abstraction of content

Comparability, equality, equivalence

UniDM

What are changes?

Deltas, operations

Classification:
simple, complex,
meaningful
changes

EUP

How to store deltas?

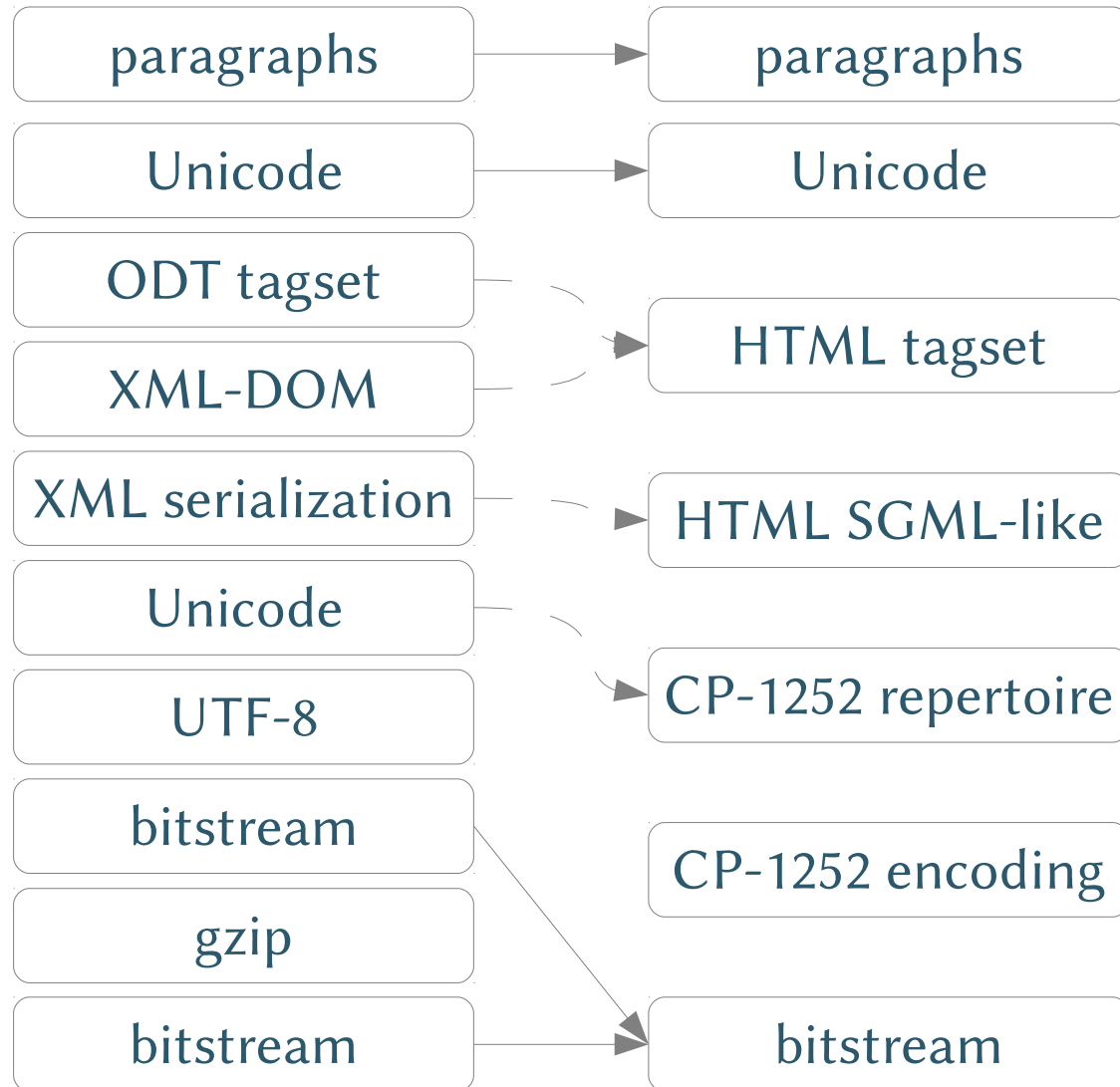
Serialization
format for
UniDM deltas.

CMV+P document model

ODT

HTML

For each level...

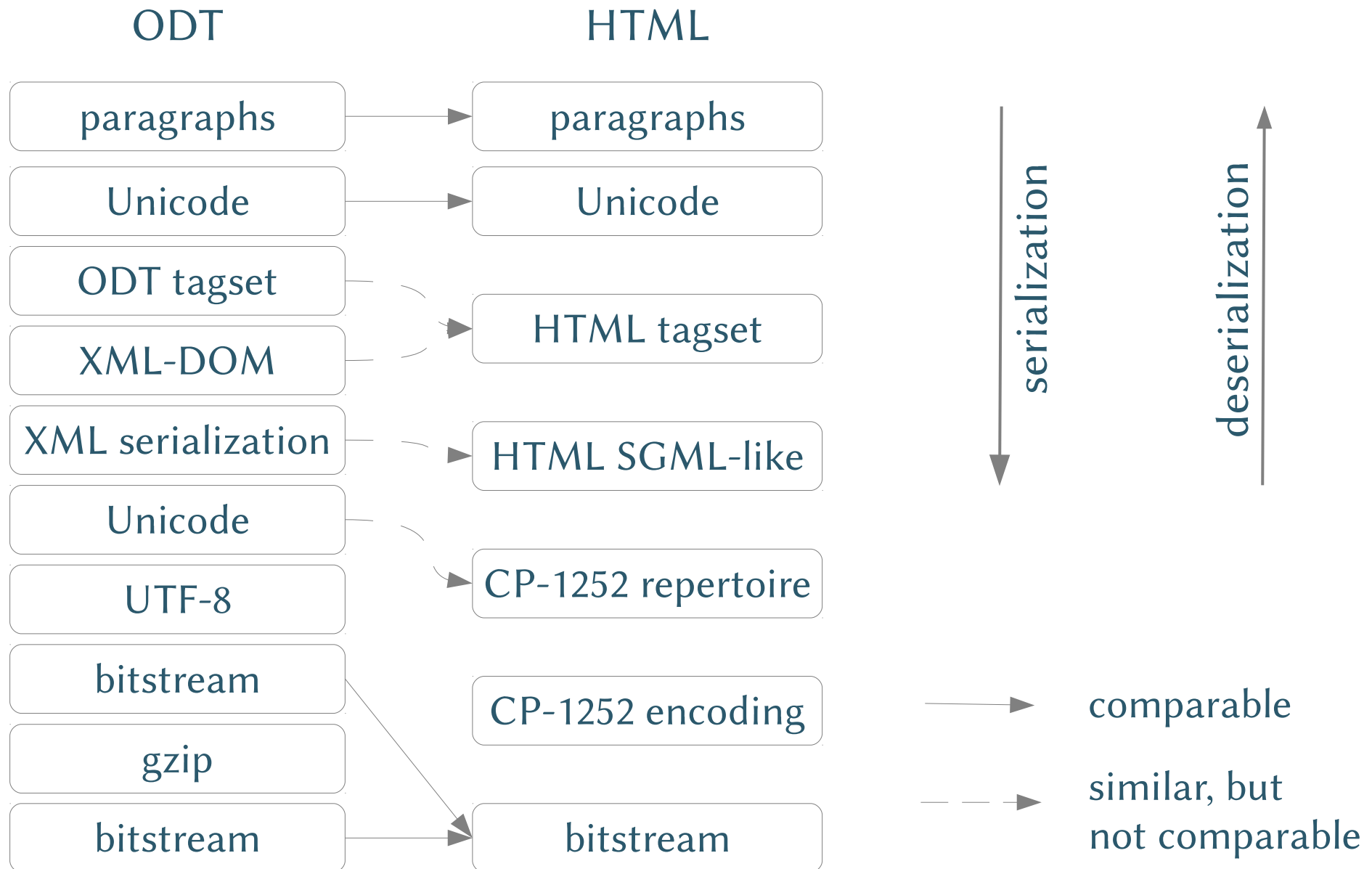


Content
typed elements

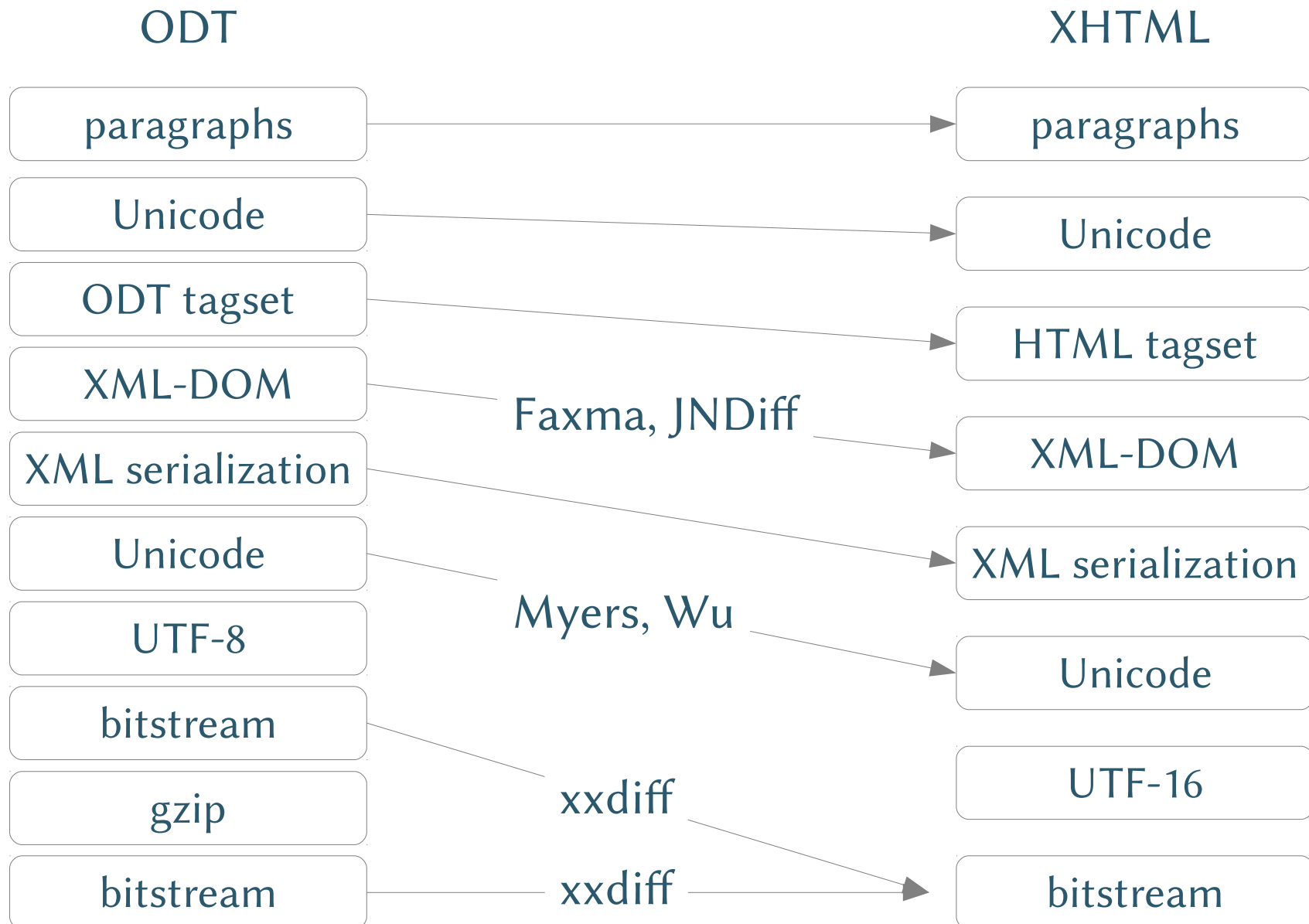
Model
set of rules, states
which kind of
content can be stored

Variants
equivalent pieces of
cnt. (for equivalence)

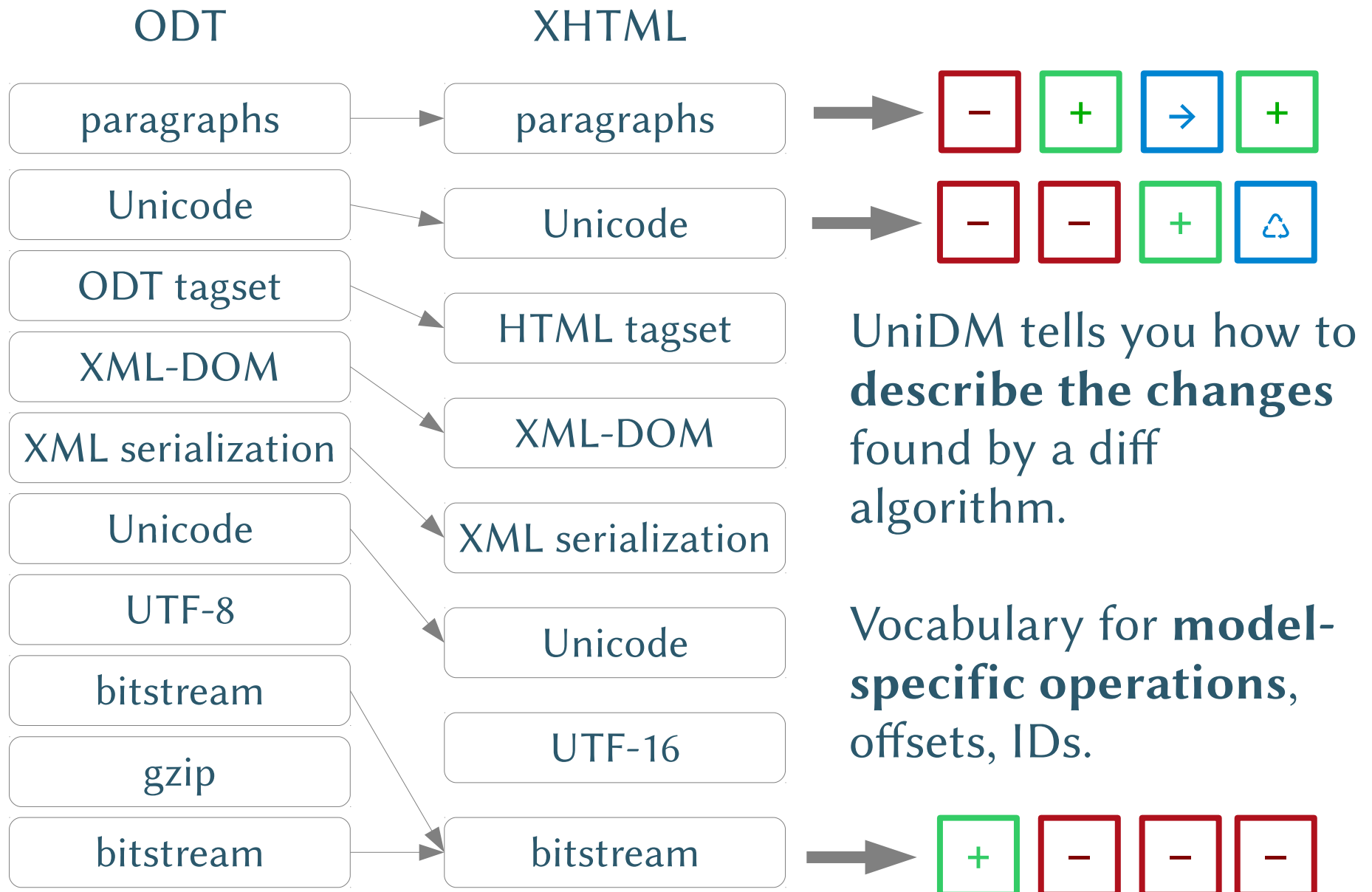
CMV+P document model



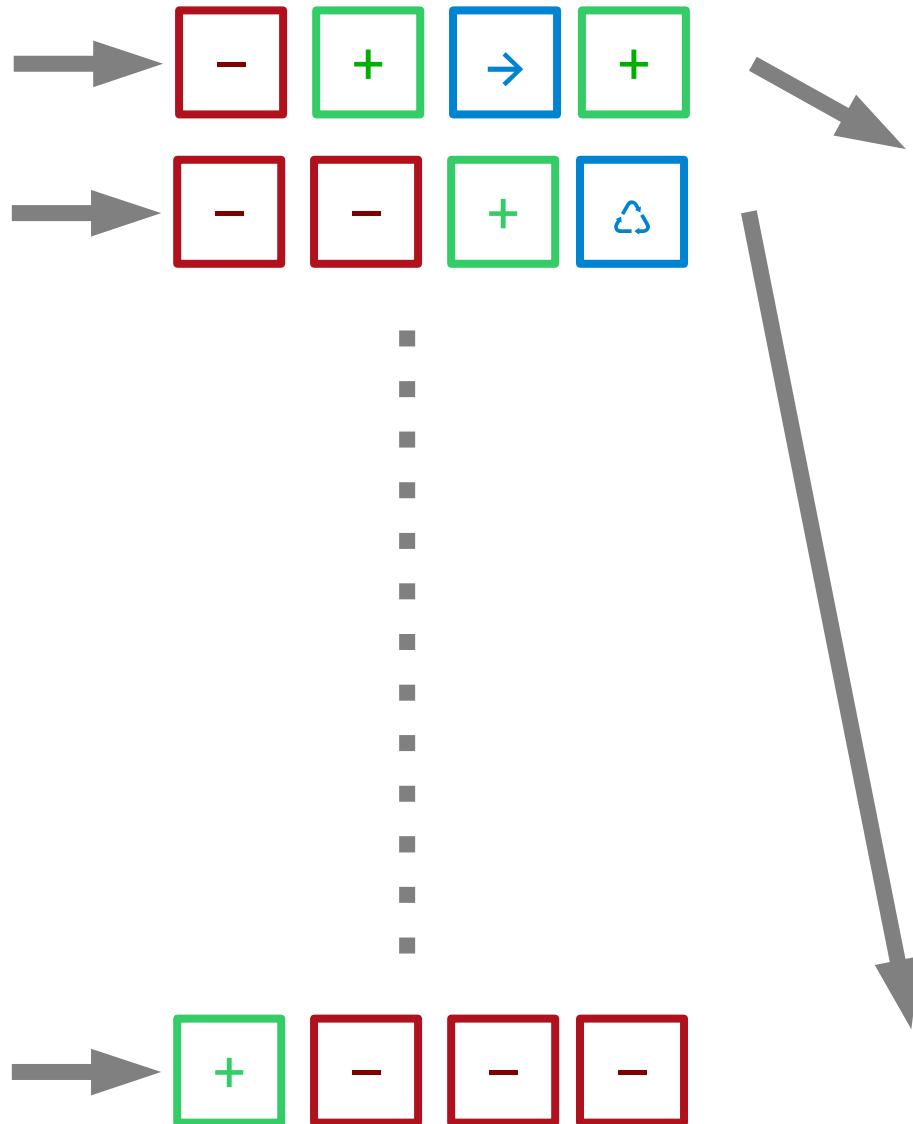
CMV+P document model



Universal Delta Model



Extended Unified Patch format



```
--- file.odt
+++ file.html
=== paragraphs / 4.21 ===
@@ -1,1 +1,4 @@@
# In the middle of the summer...
1
# -Álvaro!, cried her...
-2
# The sun was shining furiously...
3
# Paragraph "Álvaro!, cried
# her..." moved here; split into
# "Álvaro!" and "Cried her"
*Split(Move(2,3),6)
# Nobody replied...
4
=== Unicode / 1.3 ===
[...]
@@ -130,2 @@@
# (space)
0020
#-<C1> (decomposed as A + acute)
-0041
```

diffi: diff improved

Compare the content at multiple levels of abstraction, not “the file”

- Pluggable diff algorithms
- Intended as a workbench for designing diff algos
- Pre-alpha design state
- Interested? Join the diff mailing list at

<https://diffpedia.org/forum>

Thank you

Gioele Barabucci

gioele.barabucci@uni-koeln.de

