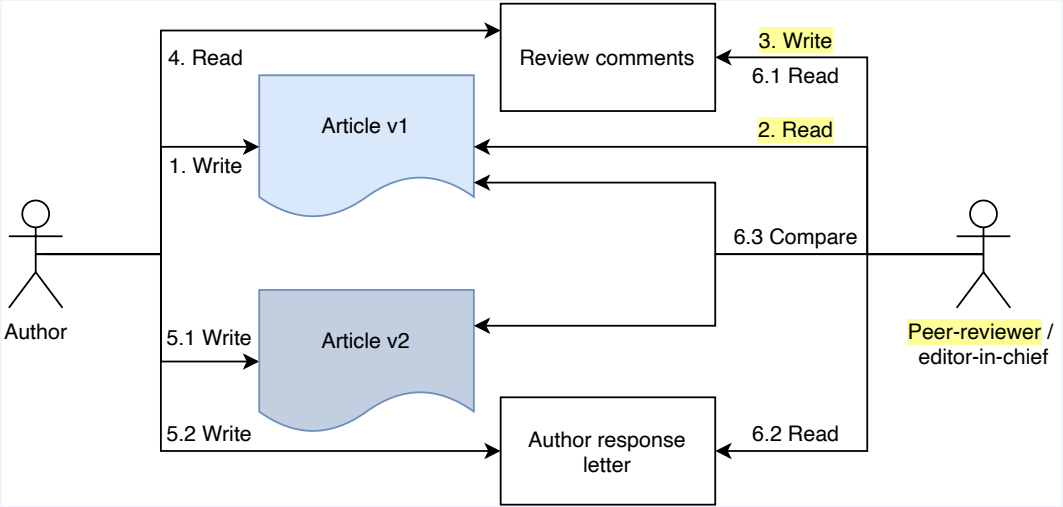
A JATS XML Comparison Algorithm for Scientific Literature

Milos Cuculovic[1,2], Frederic Fondement[1], Maxime Devanne[1], Jonathan Weber[1] and Michel Hassenforder[1]

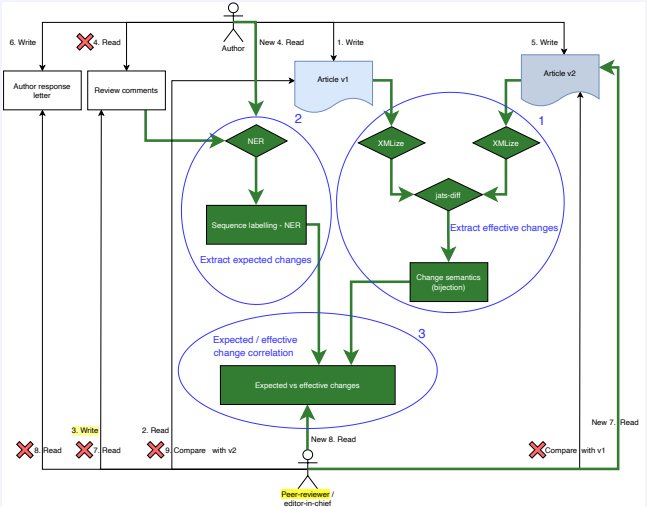
1. IRIMAS, University of Haute-Alsace, Mulhouse, France
2. MDPI, Basel, Switzerland

May 3, 2022

Publishing process



Assisted final decision making



Introduction

Comparing different versions of JATS documents is valuable for:

- Editor-in-chief
- Reviewer
- Preprint publisher
- etc.

Academic articles are available in TEX, DOCX, ODT, PDF, JATS, HTML

Advantage of the XML version:

- JATS as the de-facto standard for academic articles
- Content specific with no layout information

Research Question

Are generic XML diff algorithms suitable for JATS XML comparison?
If not, why a JATS-specific XML diff algorithm is needed?

Existing XML diff edit detection

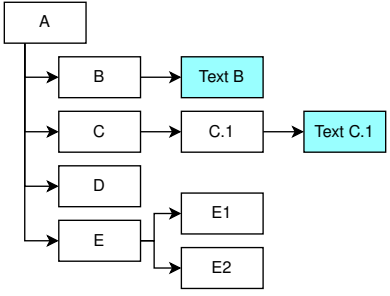
Detected edit actions

Text-edits:

- Delete
- Insert
- Update

Tree-edits:

- Delete
- Insert
- Update attribute
- Move



Existing XML diff edit detection

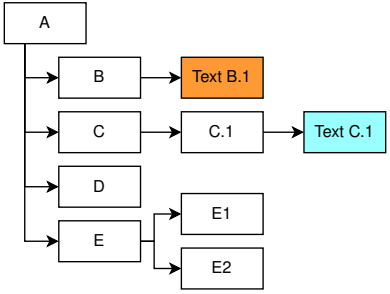
Existing edit actions

Text-edits:

- **Delete**
- **Insert**
- **Update**

Tree-edits:

- Delete
- Insert
- Update attribute
- Move



Existing XML diff edit detection

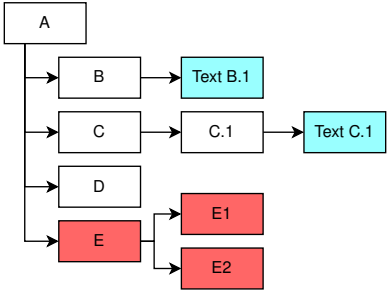
Existing edit actions

Text-edits:

- Delete
- Insert
- Update

Tree-edits:

- **Delete**
- Insert
- Update attribute
- Move



Existing XML diff edit detection

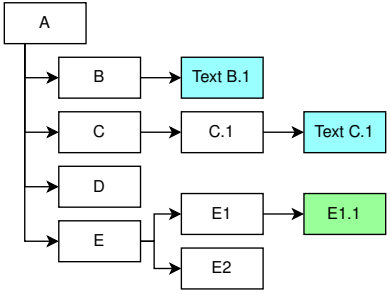
Existing edit actions

Text-edits:

- Delete
- Insert
- Update

Tree-edits:

- Delete
- **Insert**
- Update attribute
- Move



Existing XML diff edit detection

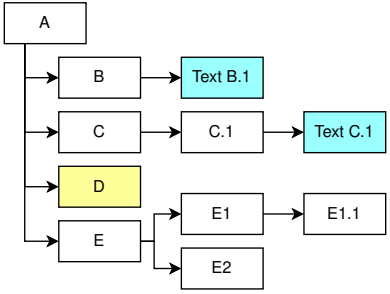
Existing edit actions

Text-edits:

- Delete
- Insert
- Update

Tree-edits:

- Delete
- Insert
- **Update attribute**
- Move



Existing XML diff edit detection

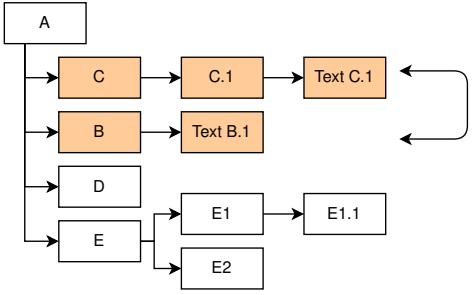
Existing edit actions

Text-edits:

- Delete
- Insert
- Update

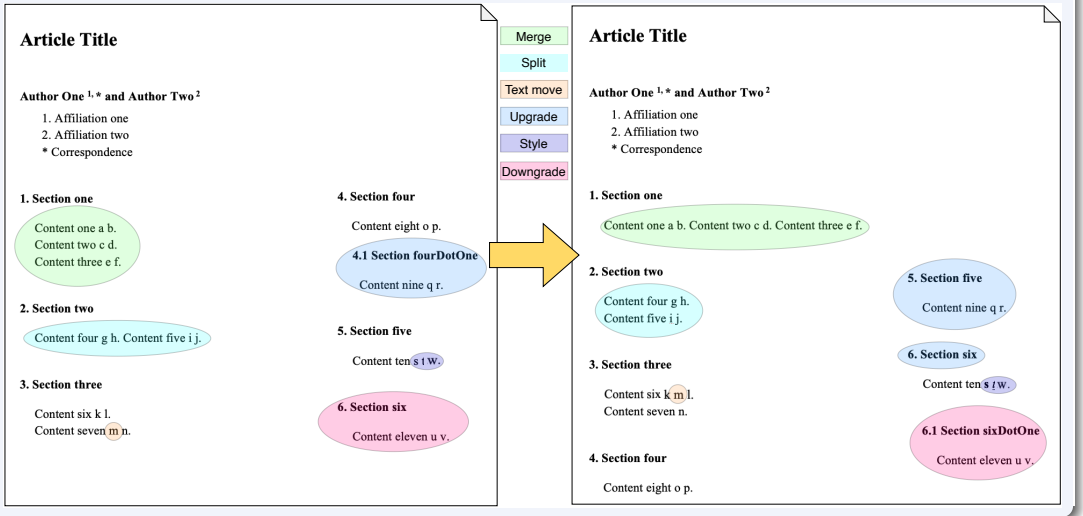
Tree-edits:

- Delete
- Insert
- Update attribute
- **Move**



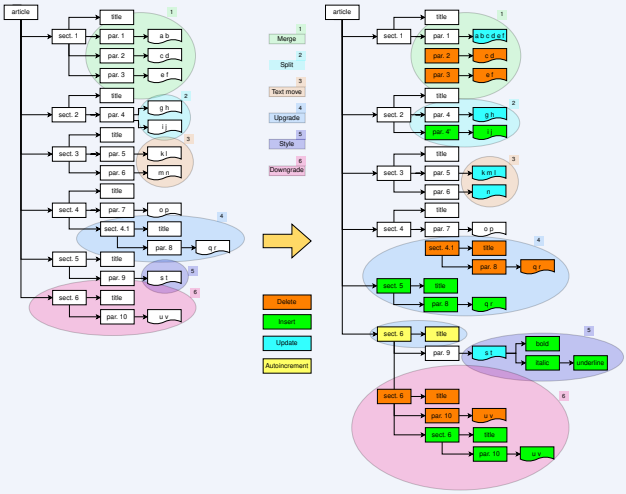
JATS comparison example

Two article versions side by side written in a text processor



JATS comparison example

Two JATS versions side by sid



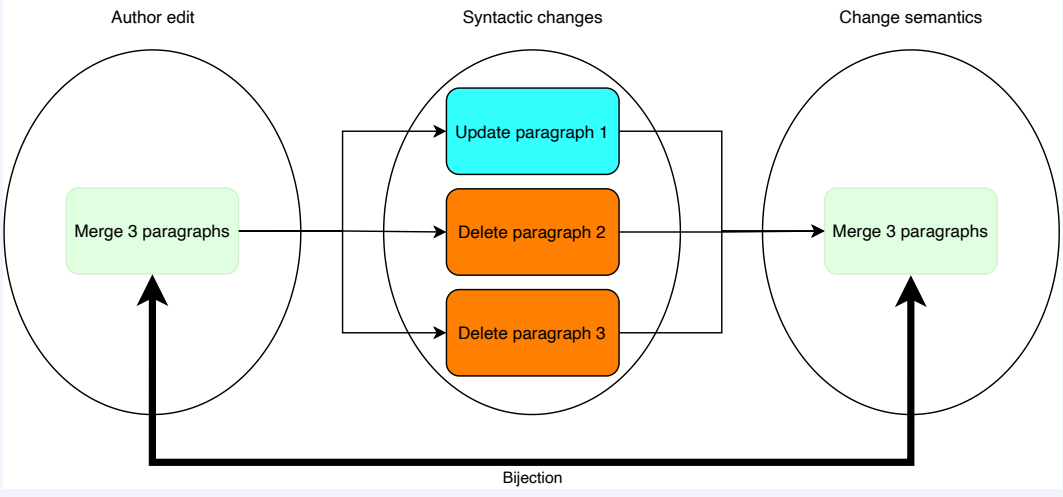
Existing XML diff algorithms evaluation

Delta output analysis

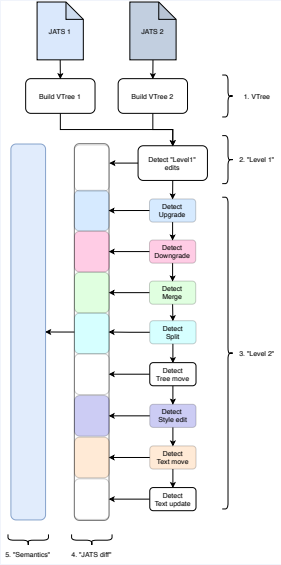
	Text edits						Tree edits										
Algorithm	Del	Ins	Mv	Upd	Style	Complex	Del	Ins	Attr	Mv	Merge	Split	Upgr	Dwngr	Style	Complex	Res/36
XOP																	0
DiffXml	1	1															2
FC-XmlDiff								2	1								3
DiffMK		1		1				2									4
Xdiff							2		2							2	6
Node-delta	1	1					2		1		1	1	1			2	10
DeltaXML	2	2					2	2				1	1	1			11
jXyDiff	2	2		1	1		1	2	2	2							13
XCC	2	2		1	1			2	2		1	1				2	14
XMLDiff	2	2		2	2	4		1					1			1	15
XyDiff	2			1	1		2	2	2	2	1	1	1				15
JNDiff	2	2		1	1	4	2	1	2	2	1	1	1	1		4	25

Purpose of jats-diff

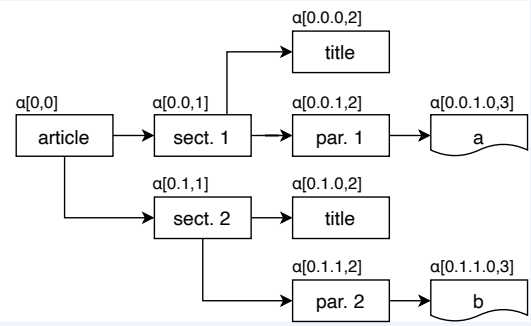
Bijection



jats-diff workflow

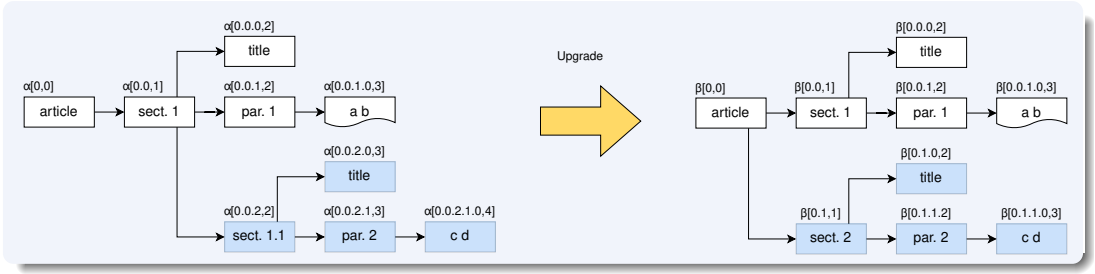


XML tree annotation



- Two JATS documents A and B
- Document A: nodes $\alpha[m, n]$ with identifier m and depth n
- Document B: $\beta[p, q]$ with identifier p and depth q

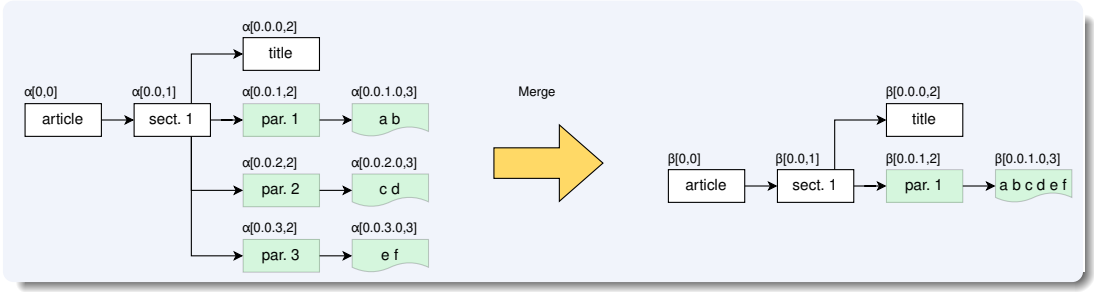
Subsection upgrade / Section downgrade



Mathematical formula

$$(\forall c\beta)(\forall c\alpha)\beta[p,q].content \simeq \alpha[m,n].content \wedge q < n \quad (1)$$

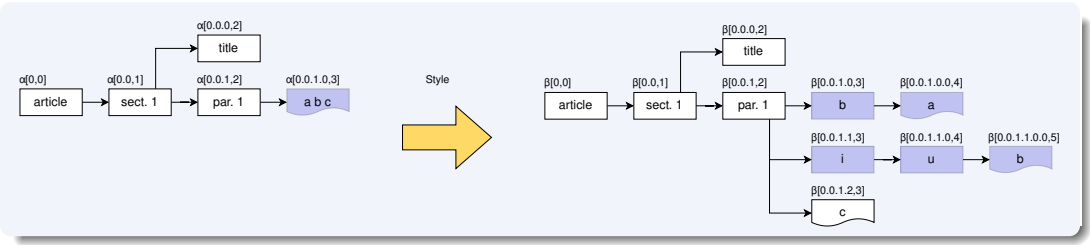
Paragraph merge and split



Mathematical formula

$$(\forall c\beta)(\forall c\alpha)\alpha[m,n].content \subset \beta[m,n].content \wedge q = n \quad (2)$$

Style edit

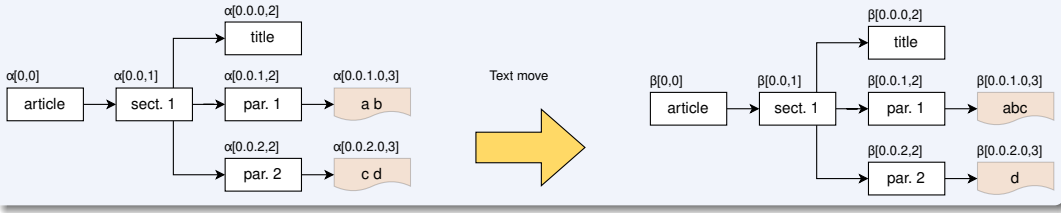


Approach

Converting style nodes to simple text using encryption:

- a converted to pure text: $\_|b|\_a\_|/b|\_$
- $<i><u>b</u></i>$ converted to pure text: $\_|i|\_|u|\_b\_|/u|\_|/i|\_$

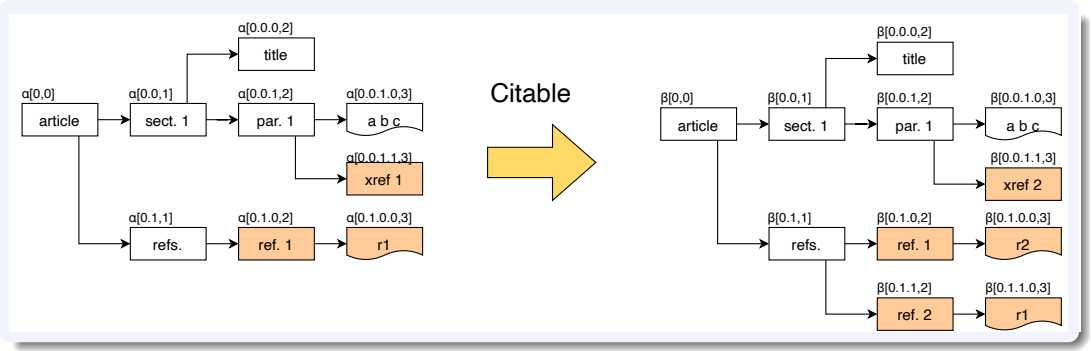
Text move



Mathematical equation

$$(\forall c\beta)(\forall c\alpha)c\beta[m,n].content - c\alpha[m,n].content \simeq c\alpha'[m,n].content - c\beta'[m,n].content \quad (3)$$

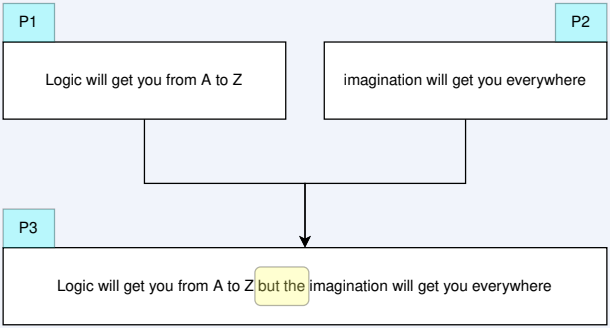
Citable objects



Approach

- $< xref >$ nodes with "ID" auto-incremental value
- Dependent on the citable object appearance order
- Scan the citable objects list and create the correspondence table

Text similarity vs. text equality



- Text similarity used while comparing text contents
- Allows to detect changes with text edit interference

Change semantics

Similarity index propagation

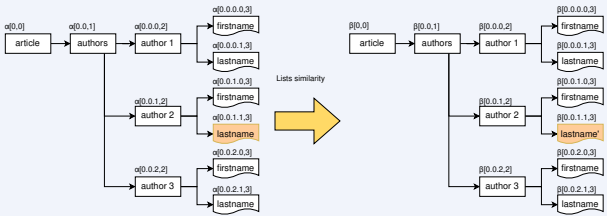
$$ParentSym = \sum_{n=0}^{n=N-1} Sn * In$$

With:

- n as the child node number
- N as the total number of child nodes for a given parent
- S as the child nodes text similarity index
- I as the child nodes text content ratio within a specific parent

Change semantics

Element lists similarity index - Authors list



Element lists similarity index

Initial: number of child elements on document A;

Final: number of child elements on document B;

Modified: number of modified child elements;

jats-diff usage - delta.xml

DTD - delta.xml

- $<delete>$
- $<insert>$
- $<update-attribute>$
- $<upgrade>$
- $<downgrade>$
- $<merge>$
- $<split>$
- $<move>$
- $<text-style-insert>$
- $<text-style-delete>$
- $<text-style-update>$
- $<text-move>$
- $<text-update>$

jats-diff usage - delta.xml

Attributes

- at: the parent node number where the child node was edited
- nodenumberA: Node number in document A
- nodenumberB: Node number in document B
- direction: parameters $<from>:<to>$ for move, split and merge
- op: type of edit for edits composed of multiple XML operation tags
- pos: for text edits, specifies the position where the edit takes place
- length: for text update, specifies the length of the modified text

Subsection upgrade

```
<upgrade at="388" nodecount="5" nodenumberB="601" op="upgradedTo" pos="5">
    <sec id="sec6"/>
</upgrade>
<upgrade nodecount="5" nodenumberA="447" op="upgradedFrom">
    <sec id="sec2dot3"/>
</upgrade>
```

Paragraph merge

```
<merge at="0" direction="9:9" nodenumberB="9" op="mergedTo" pos="4">
        <p>Text paragraph four with some additional text of paragraph five</p>
</merge>
<merge direction="9:9" nodenumberA="9" op="mergedFrom">
    <p>Text paragraph four</p>
</merge>
<merge direction="11:9" nodenumberA="11" op="mergedFrom">
    <p>with some additional text of paragraph five</p>
</merge>
```

Style edit

```
<text-style-update nodenumberB="117" op="update-style-to" pos="14">
    <i>monitoring</i>
</text-style-update>
<text-style-update nodenumberA="117" op="update-style-from" pos="14">
    <b>monitoring</b>
</text-style-update>
```

Bibliography change

```
0 - article
    * depth: 0
    * similar-word: 99.9

      606 - back
      * depth: 1
      * similar-word: 99.8

        1932 - ref-list
        * Initial: 156
        * Final: 157
        * depth: 2
        * similar-word: 99.7

          7993 - ref
          * id :B153
          * depth: 3
          * change-type: insert
```

Similarity index

```
0 - article
* depth: 0
* similar-word: 99.5
  155 - body
  * depth: 1
  * similar-word: 99.3

    156 - sec
    * id :sec1
    * 1. Introduction
    * depth: 2
    * similar-word: 99.9

      169 - p
      * depth: 3
      * change-type: text-update-text-inserted
      * similar-word: 99.3

175 - sec
* id :sec2
* 2. Materials
* depth: 2
* similar-word: 96.2

  199 - sec
  * id :sec2dot2
  * 2.2. Secname
  * depth: 3
  * similar-word: 77.8

    202 - p
    * depth: 4
    * change-type: delete
```

Current limitations

Composed changes

- Upgrade Subs. 2.3 to Sec. 6 + move $<p>$ of the upgraded subs. to Sect. 5
- Split paragraph of Sect. 1 and move Sect. 1 after Sect. 2
- Split paragraph of Sect. 1 and update the text in the second split
- Merge two paragraphs of Sect. 1 and move Sect. 1 after Sect. 2
- Merge two paragraphs of Sect. 1 and update style range in the merged paragraph
- ...

Discussion

Further development:

- Delta and semantics output files information quality
- Additional similarity indexes to detect between sentence meaning
- Change visualisation: convert JATS to HTML and annotate the changes / similarity
- NER on reviewer comments and correlation with jats-diff output

Conclusion

- jats-diff is which is able to detect additional ”Level 2” edit patterns:
  - paragraph split and merge
  - text move
  - subsection upgrade
  - section downgrade
  - style and citable object edits
- Bijection between author modifications and changes between two JATS documents
- Similarity index between different parts of the JATS document
- Application areas:
  - peer reviewers’ and the Editors-in-Chief’s decision-making process
  - preprint publishers for comparing preprint and published article versions

Thank you !

Contact details

Milos Cuculovic
MDPI - St. Alban Anlage 66, 4052 Basel, Switzerland
University of Haute Alsace, 2 Rue des Frères Lumière, Mulhouse, France

Email: milos.cuculovic@uha.fr / cuculovic@mdpi.com
LinkedIn: www.linkedin.com/in/cuculovic
Research Gate: www.researchgate.net/profile/Milos-Cuculovic
ORCID: orcid.org/0000-0003-2154-9652

jats-diff GitHub

https://github.com/milos-cuculovic/jats-diff