

# Enriched bibliographic metadata and full-texts of Swiss articles in “Krankenpflege” and “GERONTOLOGIE CH. Praxis + Forschung”

*Andrea Moritz<sup>1\*</sup>, Beatrice Hodel<sup>1</sup>, Clemens Trautwein<sup>2</sup>*

*18.04.2024*

## **Affiliations**

*1 ZHAW Zurich University of Applied Sciences, Finance & Services, University Library*

*2 ZHB – the Central and University Library of Lucerne, on behalf of Lucerne University of Applied Sciences and Arts (HSLU)*

## **Corresponding author’s email address**

*andrea.moritz@zhaw.ch*

## **Keywords**

*bibliographic metadata, green Open Access, full-text, professional journal, Switzerland*

## **Abstract**

*This data paper describes the collection process and processing steps of bibliographic metadata of articles published in the two professional journals “Krankenpflege”<sup>1</sup> (ISSN: 0253-0465) and “GERONTOLOGIE CH. Praxis + Forschung”<sup>2</sup> by members of Swiss higher education institutions (HEI). Metadata provided by the publisher itself (Gerontologie CH.) or the database CINAHL Ultimate (“Krankenpflege”) were obtained, enriched, partly checked, and corrected to create a list of articles in tabular form as Excel and CSV of all articles published by Swiss university members between 2020 and 2023. In addition, all full-texts of articles listed for the journal “GERONTOLOGIE CH. Praxis + Forschung” have been extracted from the issues and are available in PDF. The lists include all bibliographic metadata elements, facilitating comparison with existing records in Swiss repositories and identifying articles not yet indexed. Missing articles can be stored within a Swiss university repository if the author is affiliated with the respective institution.*



This work is licensed under CC BY 4.0. To view a copy of this license, visit:

<https://creativecommons.org/licenses/by/4.0/>

DOI: <https://doi.org/10.5281/zenodo.10991510>

---

<sup>1</sup> The website of the journal is <https://sbk-asi.ch/de/mitglieder/gemeinsam-stark/fachzeitschrift>. The title of the journal in French is “Soins infirmiers” (<https://sbk-asi.ch/fr/membres/forts-ensemble/revue-soins-infirmiers/>), in Italian “Cure infermieristiche”.

<sup>2</sup> The website of the journal is <https://www.gerontologie.ch/wissen/magazin>. The title of the journal in French is “GERONTOLOGIE CH. Pratique + Recherche” (<https://www.gerontologie.ch/fr/expertise/magazine>).

## Content

1	Value of the data .....	3
2	Background.....	4
3	Data description .....	5
4	Acquisition and processing of data for publication.....	7
4.1	Metadata and full-texts of articles in “GERONTOLOGIE CH. Praxis + Forschung” .....	7
4.1.1	Edits in publisher’s publication list with OpenRefine .....	8
4.1.2	Edits with Python script .....	11
4.2	Metadata and links to full texts of articles in “Krankenpflege” .....	12
4.2.1	Edits on results as HTM in OpenRefine.....	15
4.2.2	Edits on results in XML in OpenRefine.....	17
4.2.3	Edits with Python script .....	18
5	Limitations .....	18
6	Legal matters .....	18
7	Credit author statement.....	19
8	Acknowledgements .....	20
9	Declaration of competing interests.....	20
10	Attachment code snippet (JavaScript) for Adobe Acrobat .....	20
11	References .....	21

## 1 Value of the data

Initiatives to promote Green Open Access, such as the services JISC Router<sup>3</sup> and DeepGreen<sup>4</sup>, have been developed to automatically deliver full-text and metadata of scholarly publications to institutional and subject-based Open Access repositories, thereby promoting the Open Access transformation. However, these services are focused on publications from large international publishers that meet high standards in digital publishing (e.g. structured bibliographic metadata in multiple formats, DOIs as persistent identifiers, unique identifiers for affiliation data, use of professional journal management software/platforms, etc.). These publishers are therefore able to provide these services with the necessary metadata in a structured way.

In contrast, our data targets two practice-oriented scholarly or professional journals, published by smaller publishers who do not provide structured metadata on their own, nor do they utilize persistent identifiers such as DOIs or unique identifiers for affiliations (such as Ringgold, ROR, etc.) or authors (ORCID).

- This data is valuable as it provides Swiss universities with a quick overview of their researchers' publication output in the two professional journals "Krankenpflege" and "GERONTOLOGIE CH. Praxis + Forschung", for which no structured bibliographic data existed. It consists of all the bibliographic metadata elements required to accurately describe an article in a repository.
- The data can be reused by repository managers at Swiss universities to enhance their repositories with full-texts of articles that have not yet been indexed. No consent of authors is required, because universities have a non-exclusive right to republish the respective metadata and full-texts in their repositories.
- The data can be reused by third parties to enrich bibliographic databases where non-peer-reviewed articles in professional journals are under-represented.

These two use cases demonstrate possible ways of standardising bibliographic metadata and the inclusion of full-text publications in Open Access repositories. The methods and procedures can serve as a template for creating similar processes for other professional journals.

---

<sup>3</sup> <https://www.jisc.ac.uk/publications-router>

<sup>4</sup> <https://info.oa-deepgreen.de/>

## 2 Background

Formal, standardized Green Open Access policies, which are easily interpretable for researchers and repository managers, serve as crucial tools in facilitating the adoption of the green road to Open Access for scholarly publications. Green Open Access Policies of publishers typically involve granting authors permission to self-archive versions of their manuscripts in repositories or on personal websites. In consequence, self-archiving in an online repository is a form of secondary publication of scholarly works after initial publication to extend the reach and impact of research within academic and broader communities.

The project GOAL (*GOAL - Unlocking the Green Open Access Potential, 2024*) aims to develop case scenarios for adding full-texts of articles to Open Access repositories. These articles originate from journals with which the project has successfully negotiated self-archiving rights. Possible workflows should lead to as little as possible manual workload for libraries and publishers as librarians or repository managers apply the authors right to self-archive a publication. Ideally, these case scenarios should function seamlessly without necessitating direct involvement from researchers, thereby streamlining the workflow, and maintaining efficiency with minimal effort required from all parties involved. For this purpose, the project is testing semi-automated workflows that process and enrich bibliographic metadata and – if possible – full-texts provided by the publishers or third-party data providers.

The project team created data that will be easily accessible to both humans and machines, because most libraries of Swiss universities of applied sciences or universities of Teacher Education are small, have limited staff to maintain their repositories and most libraries depend on external or internal service providers to implement technical changes in their repositories. These changes would be necessary if data were provided only through machine-readable interfaces. This should be sufficient since the expected data volume is small. Additionally, the volume of articles to be archived per HEI and per year is very small (mostly double-digit numbers).

In order to find initial case studies, we focused on journals with substantial publication output authored by members of universities of applied sciences and teacher's education per year. Specifically, we concentrated on the top 30 journals within our dataset (Trautwein et al., 2022) that do not offer gold Open Access (OA) options. Availability of somehow structured bibliographic metadata was a precondition. Only publishers able to provide metadata that contains affiliation data of authors and links to/ files of full-text versions are being considered for the pilot program. Publishers needed to implement a self-archiving policy that allows universities to archive full-texts of their authors without their consent.

This paper documents these specific workflows for enriching repositories according to the publisher's self-archiving policies. The result of this data creation process are bibliographic metadata and full-texts of articles after data collection from the professional journals "GERONTOLOGIE CH. Praxis + Forschung" and "Krankenpflege". The data facilitate the inclusion of all possible full-texts into the repositories of project partners and beyond. Actors in the workflow are the ZHAW university library and publishers/editors of the mentioned journals.

### 3 Data description

The final data table contains bibliographic metadata in **CSV and Excel file “bibliographic-metadata-articles-Krankenpflege-GerontologieCH”** (Moritz & Hodel, 2024) with the dimension of 450 rows and 21 columns. Its structure is described in the following table:

	<i>Column name for metadata in CSV or Excel</i>	<i>Description of the content in this column</i>	<i>Defined values</i>	<i>Mandatory (M) vs. optional (O)</i>
1	<b>affiliation normalized</b>	Affiliation of Swiss HEI author	normalized with preferred name according to swissuniversities member list by hand or OpenRefine reconciliation (wikidata) (attachment A)	M
2	<b>affiliation published</b>	Affiliation given in the article or provided by the publisher	Free text	O
3	<b>dc.contributor.author</b>	Authors	Pattern “LastName, FirstName”; Delimiter between authors is «     »	M
4	<b>dc.title</b>	Title	Free text	M
5	<b>Issue</b>	Issue	Free text	O
6	<b>Volume</b>	Volume	Free text	O
7	<b>dc.date.issued</b>	Publication year	Pattern XXXX	M
8	<b>dc.relation.ispartof</b>	Journal Title	Free text	M
9	<b>pages.start</b>	Pages (start)	Digit	M
10	<b>pages.end</b>	Page (end)	Digit	M
11	<b>dc.identifier.issn</b>	ISSN	Pattern XXXX-XXXX	M
12	<b>dc.publisher</b>	Publisher	Delimiter between publishers is «     »	M
13	<b>dc.identifier.uri</b>	URL of issue/article	Free text	O
14	<b>dc.language.iso</b>	Language code	ISO 639-1, Code for French, German, Italian: “fr”, “de”, “it”	O
15	<b>publication.status</b>	Publication status allowed for self-archiving	Possible values are: “publishedVersion”; “acceptedVersion”; “submittedVersion”	M
16	<b>dc.subject</b>	Keywords	Free text, Provided by the publisher	O
17	<b>dc.description.abstract</b>	Abstract	Free text, Delimiter between subjects is «     »	O
18	<b>dc.rights</b>	Rights	“licence according to publisher”	M
19	<b>dcterms.type</b>	Material	“text”	M
20	<b>note</b>	Notes (e.g. in case of multilingual journals and translations)	Free text	O
21	<b>file.name</b>	file name of full-text in ZIP-file (“Gerontologie”)		O

**Table 1:** column names in CSV and Excel file “bibliographic-metadata-articles-Krankenpflege-GerontologieCH”

Each metadata element is represented by a column. Necessary data elements to describe an article properly are – wherever possible – in line with the metadata terms of Dublin Core<sup>5</sup>.

<sup>5</sup> <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>

The CSV/Excel contains all articles published in “GERONTOLOGIE CH. Praxis + Forschung” (2020-2023, except last issue, total amount: 126 articles) and “Krankenpflege” (2020-2023, total amount: 307 articles) that were written by a contributor from a Swiss Higher Education institution. If an article was written by members of different Swiss universities, it is repeated in the table. These articles differ only in the value of the column “affiliation normalized”. Total number of rows (articles) is 449 (including repeated articles).

The **ZIP-file “fulltexts-GerontologieCH-2020-2023.zip”** (Moritz, 2024a) includes all full-texts of articles mentioned in the CSV/Excel for the journal “GERONTOLOGIE CH. Praxis + Forschung”. All issues provided on the publisher’s website “Gerontologie CH.” were split into single PDFs/A (one PDF per article). The file name of each PDF is created in the pattern “Lastname of first author\_first 20 characters of title<sup>6</sup>\_publication year” (example: ‘Seifert\_Digitalisierung in A\_2022.pdf’).

PDFs include the following embedded metadata fields:

- dc:creator: authors of article according to metadata provided by the publisher
- dc:description : abstract of article according to metadata provided by the publisher
- dc:title: title of article according to metadata provided by the publisher
- dc:rights: Copyright©, Gerontologie CH
- xmpRights:Marked: True
- xmpRights:WebStatement: <https://www.gerontologie.ch/impressum>

The PDF files all have the property "Show document title" when opening the file.

The **CSV and Excel File “swissuniversities-member-list-2023”** (Moritz & Hodel, 2024) contains all the names of official members of swissuniversities (swissuniversities, 2023). The table was used as reference point to normalize affiliation data. For further reuse, the member list has been enhanced with information about Wikidata QID<sup>7</sup>, Gemeinsame Normdatei (GND)<sup>8</sup> and Research Organization Registry (ROR) ID<sup>9</sup>. The Wikidata, GND and ROR IDs of the institutions in question were researched and supplemented by hand to be able to use the name of the swissuniversities’ list as a default value in automated enrichment processes (column "affiliation normalized" in CSV and Excel file “bibliographic-metadata-articles-Krankenpflege-GerontologieCH”). These IDs form the basis for the automated comparison of the published affiliations with the corresponding names of the universities in these standards data.

---

<sup>6</sup> Invalid characters for file names in Windows and Linux have been removed.

<sup>7</sup> <https://www.wikidata.org>

<sup>8</sup> [https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd\\_node.html](https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html)

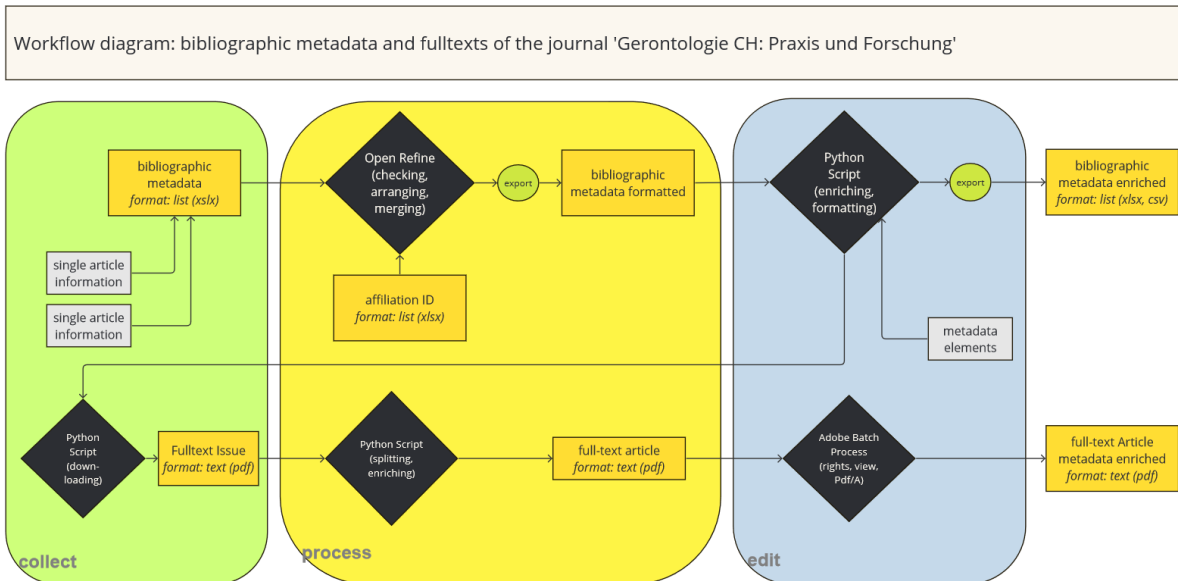
<sup>9</sup> <https://ror.org/>

## 4 Acquisition and processing of data for publication

This chapter describes the acquisition and processing of raw data to produce the final merged dataset from the two pilot journals.

### 4.1 Metadata and full-texts of articles in “GERONTOLOGIE CH. Praxis + Forschung”

Processing the bibliographic metadata for “GERONTOLOGIE CH. Praxis + Forschung” involved a series of steps, starting from data collection, followed by processing, editing and transformation. These steps are visualized in the following flowchart:



The publisher has provided an Excel file with all articles published by Swiss university members from 2020 to 2023 (issue 1 and 2). The Excel file was filled in by copying all the necessary information from the online version of the issues in the PDF and pasting it into the appropriate columns by hand by a representative of the publisher.

The Excel file consists of the following columns with bibliographic metadata for articles in both published versions (German and translated French version, and vice versa). These two language versions are published together in the print version of the journal. However, in the electronic version (PDF), they are published separately.

- Band
- Heft
- dt. URL des Hefts
- dt. Titel
- fr. Titel
- dt. Zusammenfassung
- fr. Zusammenfassung
- Autoren
- Seitenzahl (Beginn)
- Seitenzahl (Ende)
- Hochschule (dt. Version)

#### 4.1.1 Edits in publisher's publication list with OpenRefine

This Excel was checked and further preprocessed with OpenRefine (<https://openrefine.org/>, Version 3.7.8)

1. Trimming white spaces at the beginning or end of text
2. Reorder Columns
3. Create a new column with a different order of names in accordance with the desired pattern ***LastName, FirstName | LastName, FirstName*** based on column "Autoren" (functionality: split column based on this column and join columns in new order with new delimiter)

JSON:

```
[
  {
    "op": "core/column-addition",
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "baseColumnName": "Autoren",
    "expression": "grel:value",
    "onError": "set-to-blank",
    "newColumnName": "Autoren sortiert",
    "columnInsertIndex": 2,
    "description": "Create column Autoren sortiert at index 2 based on column Autoren using expression grel:value"
  },
  {
    "op": "core/multivalued-cell-split",
    "columnName": "Autoren sortiert",
    "keyColumnName": "Hochschule (dt. Version)",
    "mode": "separator",
    "separator": ",",
    "regex": false,
    "description": "Split multi-valued cells in column Autoren sortiert"
  },
  {
    "op": "core/text-transform",
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "columnName": "Autoren sortiert",
    "expression": "value.trim()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10,
    "description": "Text transform on cells in column Autoren sortiert using expression value.trim()"
  },
  {
    "op": "core/column-split",
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "columnName": "Autoren sortiert",
    "guessCellType": true,
    "removeOriginalColumn": true,
    "mode": "separator",
    "separator": " ",
    "regex": false,
    "maxColumns": 2,
    "description": "Split column Autoren sortiert by separator"
  },
  {
    "op": "core/text-transform",
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "columnName": "Autoren sortiert 2",
```



```

      "expression": "join ([coalesce(cells['Autoren sortiert 2'].value, ''), coalesce(cells['Autoren sortiert 1'].value, '')), ', ')",
      "onError": "keep-original",
      "repeat": false,
      "repeatCount": 10,
      "description": "Text transform on cells in column Autoren sortiert 2 using expression join ([coalesce(cells['Autoren sortiert 2'].value, ''), coalesce(cells['Autoren sortiert 1'].value, '')), ', ')"
    },
    {
      "op": "core/column-removal",
      "columnName": "Autoren sortiert 1",
      "description": "Remove column Autoren sortiert 1"
    },
    {
      "op": "core/multivalued-cell-join",
      "columnName": "Autoren sortiert 2",
      "keyColumnName": "Hochschule (dt. Version)",
      "separator": "||",
      "description": "Join multi-valued cells in column Autoren sortiert 2"
    }
  ]
]

```

4. Splitting articles with the participation of authors from different universities into separate lines (one line per different university)

JSON:

```

[
  {
    "op": "core/multivalued-cell-split",
    "columnName": "Hochschule (dt. Version)",
    "keyColumnName": "Hochschule (dt. Version)",
    "mode": "separator",
    "separator": ",",
    "regex": false,
    "description": "Split multi-valued cells in column Hochschule (dt. Version)"
  },
  {
    "op": "core/text-transform",
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "columnName": "Hochschule (dt. Version)",
    "expression": "value.trim()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10,
    "description": "Text transform on cells in column Hochschule (dt. Version) using expression value.trim()"
  }
]

```

5. Fill down empty cells

6. Reconciliation of mentioned university in column “Hochschule (dt. Version)” with Wikidata

JSON:

```

[
  {
    "op": "core/column-addition",
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "baseColumnName": "Hochschule (dt. Version)",
    "expression": "grel:value",
    "onError": "set-to-blank",
    "newColumnName": "QID",
    "columnInsertIndex": 1,
    "description": "Create column QID at index 1 based on column Hochschule (dt. Version) using expression grel:value"
  },
  {
    "op": "core/recon",
    "engineConfig": {

```

```

    "facets": [],
    "mode": "row-based"
  },
  "columnName": "QID",
  "config": {
    "mode": "standard-service",
    "service": "https://wikidata.reconci.link/de/api",
    "identifierSpace": "http://www.wikidata.org/entity/",
    "schemaSpace": "http://www.wikidata.org/prop/direct/",
    "type": {
      "id": "Q38723",
      "name": "Hochschule"
    }
  },
  "autoMatch": true,
  "columnDetails": [],
  "limit": 0
},
"description": "Reconcile cells in column QID to type Q38723"
}
]

```

7. Manually reconcile all entries not matched automatically.

8. Obtain QID of reconciled Wikidata entry.

JSON:

```

[
  {
    "op": "core/column-addition",
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "baseColumnName": "QID",
    "expression": "cell.recon.match.id",
    "onError": "set-to-blank",
    "newColumnName": "QID value",
    "columnInsertIndex": 2,
    "description": "Create column QID value at index 2 based on column QID using expression cell.recon.match.id"
  }
]

```

9. Cross Check the QID obtained in this way with the swissuniversities membership list (Moritz & Hodel, 2024) for uniform normalization of affiliation in a new "affiliation normalized" column.

JSON:

```

[
  {
    "op": "core/column-addition",
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "baseColumnName": "QID value",
    "expression": "grel:cell.cross(\n      \"swu member list 2023\", \n      \"QID (wikidata)\" )\n .cells[\"preferred name (long)\"]\n .value[0]",
    "onError": "set-to-blank",
    "newColumnName": "affiliation normalized",
    "columnInsertIndex": 3,
    "description": "Create column affiliation normalized at index 3 based on column QID value using expression grel:cell.cross(\n      \"swu member list 2023\", \n      \"QID (wikidata)\" )\n .cells[\"preferred name (long)\"]\n .value[0]"
  }
]

```

10. Manual verification of author names

11. Export as Excel for further processing with Python

#### 4.1.2 Edits with Python script

To minimize manual effort on the publisher’s end, only the essential fields were manually added in the original Excel file. Any metadata element that could be calculated unattended was added using a Python script (Moritz, 2024). Missing static metadata elements (such as publisher, ISSN, link to French version etc.) were added automatically.

All columns of this Excel were mapped to the GOAL metadata schema with the help of a Python script.

	Metadata element in CSV	Column in delivered excel by publisher	Metadata element added by GOAL project (x)
1	affiliation normalized		x
2	affiliation published	Hochschule (dt. Version)	
3	dc.contributor.author	Autoren	
4	dc.title	dt. Titel fr. Titel	
5	issue	Heft	
6	volume	Band	
7	dc.date.issued		x
8	dc.relation.ispartof		x
9	Pages.start	Seitenzahl (Beginn)	
10	pages.end	Seitenzahl (Ende)	
11	dc.identifier.issn		x
12	dc.publisher		x
13	dc.identifier.uri	dt. URL des Hefts	
14	dc.language.iso		x
15	publication.status		x
16	dc.subject		x, missing values are coded as "None"
17	dc.description.abstract	dt. Zusammenfassung fr. Zusammenfassung	
18	dc.rights		x
19	dcterms.type		x
20	note		x
21	file.name		x

**Table 2:** Merge between GOAL table schema, delivered metadata from publisher and added data by GOAL.

The connection between the translated versions (German and French) in the original article list provided by the publisher is maintained via a note in the column “note”. In the provided CSV/Excel file every language version of an article is in a separate row.

The data records were enriched under the following conditions:

- Journal is published bilingually
- Author composition, page numbering, volume and issue details for article versions in German and French are the same (with rare exceptions regarding page numbering)
- Articles appear in separate electronic editions
- Download of the issues (one PDF file per issue) is freely available from the publisher’s website  
 French Version: <https://www.gerontologie.ch/fr/expertise/magazine>  
 German Version: <https://www.gerontologie.ch/wissen/magazin>

The URLs of the full text files of the magazines differ only in the language code (regex "\_de" for German, "\_fr" for French).

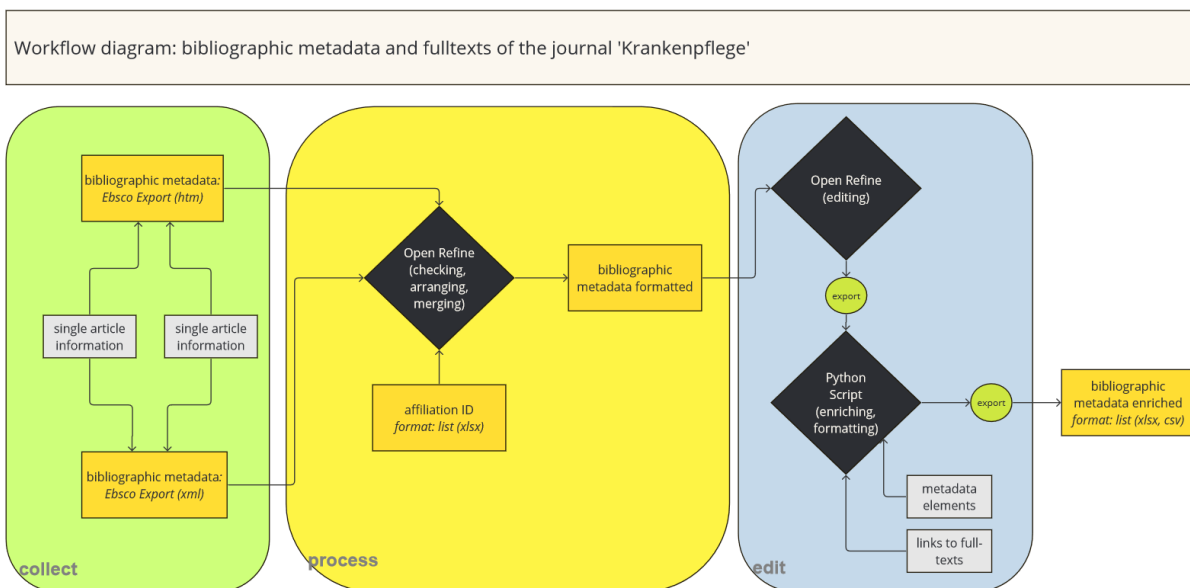
Additionally, all issues provided on the publisher’s website were downloaded and split into single PDFs (one PDF per article) based on metadata (page numbers) provided by the publisher. PDFs are added with the following metadata fields with the same Python script:

- dc:creator: authors of article according to metadata provided by the publisher
- dc:description : abstract of article according to metadata provided by the publisher
- dc:title: title of article according to metadata provided by the publisher
- dc:rights: Copyright©, Gerontologie CH
- xmpRights:Marked: True
- xmpRights:WebStatement: <https://www.gerontologie.ch/impressum>

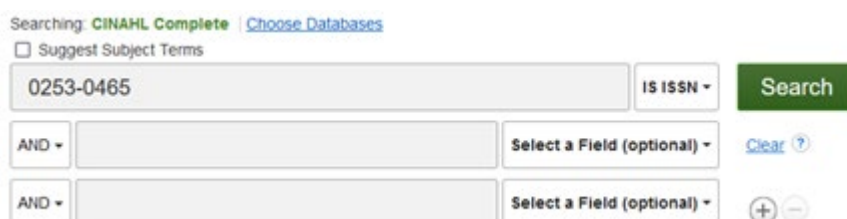
Writing of XMP-Rights and dc:rights in the metadata, saving as PDF/A, and adjustments to the view of the PDFs were done with the help of a customized Adobe batch process including a JavaScript code (see Attachments 3.2.).

#### 4.2 Metadata and links to full texts of articles in “Krankenpflege”

Processing the bibliographic metadata for “Krankenpflege” followed a pathway from data collection, processing, editing and transformation. These steps are visualized in the following flowchart:



For “Krankenpflege”, ZHAW searched all articles indexed in the database CINAHL Ultimate (EBSCO, <https://www.ebsco.com/de-de/produkte/datenbanken/cinahl-ultimate>) from 2020 to 17.10.2023 (see screenshot below). And from 18.10.2023 to 31.12.2023. The ISSN of the journal was used for the search query. All search results were exported as XML file (as compressed zip file) and HTM files and further processed by the ZHAW.



The structure of the XML looked like this:

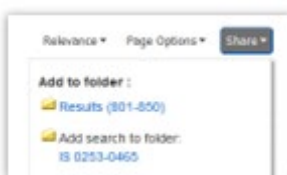
```
<records>
<rec resultID="1">
<header shortDbName="ccm" longDbName="CINAHL Complete" uiTerm="172963833">
<controlInfo>
<bkinfo />
<dissinfo />
<jinfo>
<jtl>Krankenpflege: Soins Infirmiers</jtl>
<issn>02530465</issn>
</jinfo>
<pubinfo>
<dt year="2023" month="10" day="01">2023</dt>
<iid>10</iid>
</pubinfo>
<artinfo>
<ppf>76</ppf>
<ppct>4</ppct>
<formats />
<tig>
<atl>«In primo piano non è la restrizione, ma il modo in cui gestirla.».</atl>
</tig>
<aug>
<au>Pfammatter, Danielle</au>
<au>Schwager, Christa</au>
<affil>Infermiera esperta in riabilitazione CSP, MAS Rehabilitation Care, formatrice in cinestetica S3.</affil>
</aug>
<sug />
<pubtype>Periodical</pubtype>
<doctype>Journal Article</doctype>
</artinfo>
<language>Italian</language>
</controlInfo>
<displayInfo>
<plink>
<url>https://search.ebscohost.com/login.aspx?direct=true&db=ccm&AN=172963833&site=ehost-live</url>
</plink>
</displayInfo>
</header>
</rec>
```

The XML file lists only the first mentioned affiliation of authors, other affiliations are missing.

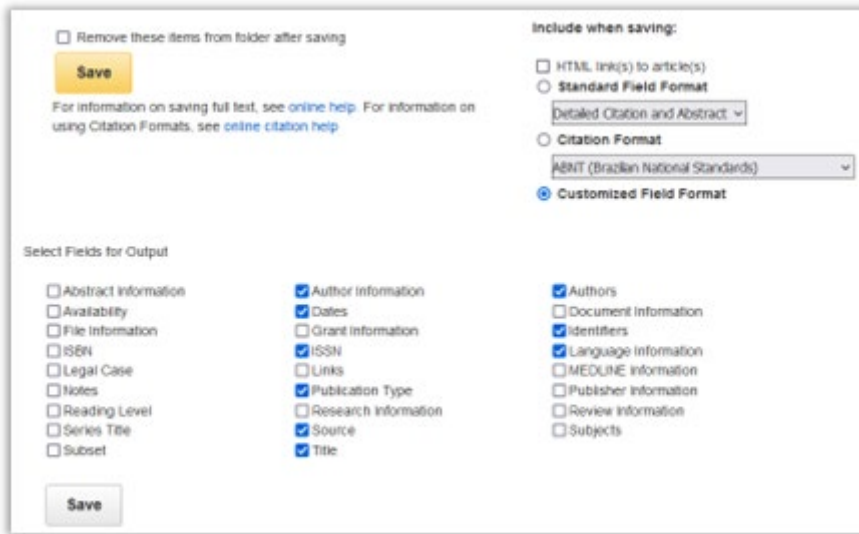
To add the missing information to the XML file, the search result for all the indexed articles of “Krankenpflege” (tranche 1.1.2020-17.10.2024 1107 records), was exported in HTM format (packaged in 6 HTM files containing up to 200 records each) from database CINAHL Ultimate. The HTM files include all affiliations of authors.

The additional tranche of exported articles published 18.10.2023 to 31.12.2023 included 40 records. The search result from the database CINAHL Ultimate were again exported as XML and HTM.

To export the results from our query as HTM files, all items were added to folder and exported via the sharing option “Save as file”.



These were the export settings used:



Structure of the record list (HTM) in the browser looks like this:

```

Record: 4
Title:
«Beckenbodenbeschwerden sind zu keinem Zeitpunkt normal».
Authors:
Enaux, Jennifer; 1Bernet, Madeleine2
Affiliation:
1MSc, dipl. Pflegefachfrau, Leiterin Pflegeentwicklung Spitalregion Rheintal Werdenberg Sarganserland, Doktorandin in
Medizinischen Wissenschaften an der Privaten Universität im Fürstentum Liechtenstein (UFL), Mitglied Akademische
Fachgesellschaft für Frauengesundheit VFP.
2Pflegefachfrau MScN, Wissenschaftliche Mitarbeiterin und Studienleiterin an der Berner Fachhochschule Gesundheit; Co-
Präsidentin Akademische Fachgesellschaft für Frauengesundheit VFP.
Source:
Krankenpflege: Soins Infirmiers 2023; (7/8): 20-22. (3p)
Publication Type:
Journal Article
Language:
German
ISSN:
0253-0465
Entry Date:
In Process
Revision Date:
20230725
Accession Number:
164923557
Database:
CINAHL Complete
    
```

The relevant information is contained in the HTM element <div id="records">:

```

1 <div id="records">
2 <br><strong>Record: 1</strong><dl class="print
3 "citation_field_label"><strong><span class="me
Affiliation:</span></strong></dt><dd data-auto
"citation_field_label"><strong><span class="me
Type:</span></strong></dt><dd data-auto="citat
</span></dd><dt data-auto="citation_field_labe
</span></strong></dt><dd data-auto="citation_f
    
```

The contents of the <div id="records"> of all seven files were copied into one file, which was then further processed using OpenRefine

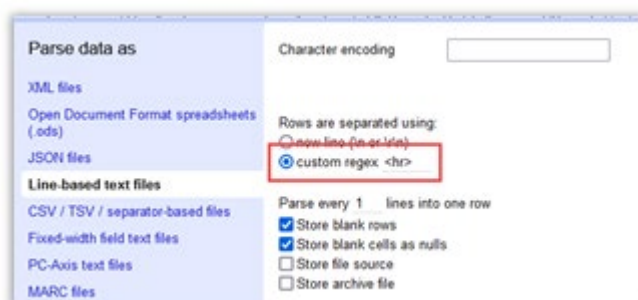
The HTM files and XML file of articles of “Krankenpflege” were pre-processed using Open Refine (merging records in HTM files with records in XML file adding all affiliations of authors and page numbers; normalization, and clustering of affiliation data on organizational level by hand; alignment of data to the structure described below).

XML elements are mapped with the GOAL metadata structure:

Meta data element in CSV	XML-Tags	Note
<b>Affiliation normalized</b>	None	done manually
<b>Affiliation published</b>	rec - header - controllInfo - aug - affil	
<b>dc.contributor.author</b>	rec - header - controllInfo - aug - au	
<b>dc.title</b>	rec - header - controllInfo - artinfo - tig - atl	
<b>issue</b>	rec - header - controllInfo - pubinfo – iid	
<b>volume</b>	rec - header - controllInfo - pubinfo - dt	
<b>dc.date.issued</b>	rec - header - controllInfo - pubinfo – dt	Publication year is the same as volume
<b>dc.relation.ispartof</b>	rec - header - controllInfo - jinfo – jtl	
<b>pages.start</b>	rec - header - controllInfo - artinfo – ppf	
<b>pages.end</b>	None	XML contained no end page, pages were extracted from matching HTM record (field “source”)
<b>dc.identifier.issn</b>	rec - header - controllInfo - jinfo - issn	XML contained ISSN in pattern xxxxxxxx, was transformed to xxxx-xxxx
<b>dc.publisher</b>	None	
<b>dc.identifier.uri</b>	None	
<b>dc.language.iso</b>	rec - header - controllInfo - language	Data included no ISO-codes, data was normalized by hand
<b>publication.status</b>	None	
<b>dc.subject</b>	rec - header - controllInfo - artinfo - sug - subj - subj	Only few records included keywords
<b>dc.rights</b>	None	
<b>dc.description.abstract</b>	rec - header - controllInfo - artinfo – ab	Only few records included abstracts
<b>dcterms.type</b>	None	
<b>note</b>	rec - header - displayInfo - pLink - url	

**Table 3:** Mapping of XML elements and column names in GOAL table structure.

In OpenRefine, a new project was created by copying the contents of the new file, which included the data within the <div id="records">, to the Clipboard. The data was parsed as "Line-based text files," with rows being separated using <hr>.



#### 4.2.1 Edits on results as HTM in OpenRefine

The new project contained 1107 rows/records (or 40) with one column “Column 1”. From this column the following information was extracted: authors (for reasons of clarity and comprehensibility of

following data cleansing), affiliation, end pages of publication, accession number (to merge with the other project in OpenRefine)

1. Add column “Authors” based on “Column 1”:

If author information is available, it can be found within the second dd-element. Use GREL:

```
if(value.contains("Authors:"), value.parseHtml().select("dd")[1].htmlText(), "")
```

2. Add column “Affiliation” based on “Column 1”:

If authors have an affiliation, this can be found in the third dd-element. Use GREL:

```
if(value.contains("Affiliation:"), value.parseHtml().select("dd")[2].htmlText(), "")
```

3. Edit affiliations:

The affiliations (and authors) are numbered (e.g.: 1Pflegeexperte APN, RN, MScN, Psychiatrische Universitätsklinik Zürich, Klinik für Alterspsychiatrie 2Dozentin Berner Fachhochschule). In order to receive separate rows for affiliations per author, use the transformation “value.substring(1)” to remove the initial number 1 from each affiliation, and then apply “Split multi-valued cells” with the separator “\d”.

“Affiliation” was duplicated by applying “Add column based on this column” without any transformations. This leaves “Affiliation\_copy” as a backup with the original affiliation.

Affiliation was normalized manually based on the Excel file “swissuniversities-member-list-2023” (Moritz & Hodel, 2024).

These hospitals/centers are affiliated with the following universities:

Swiss Higher Education Institution	Affiliated hospitals/centers
<b>Universität Zürich</b>	<ul style="list-style-type: none"> <li>• Universitätsspital Zürich</li> <li>• Universitäts-Kinderspital</li> <li>• Universitätsklinik Balgrist</li> <li>• Psychiatrische Universitätsklinik</li> <li>• Institut für Notfallmedizin</li> </ul>
<b>Universität Basel</b>	<ul style="list-style-type: none"> <li>• Universitätsspital Basel</li> <li>• Universitätskinderspital beider Basel</li> <li>• Universitäre Psychiatrische Kliniken</li> <li>• Universitäre Altersmedizin Felix Platter</li> <li>• Clarunis – Universitäres Bauchzentrum Basel</li> </ul>
<b>Universität Bern</b>	<ul style="list-style-type: none"> <li>• Inselspital Bern</li> </ul>
<b>Université de Genève</b>	<ul style="list-style-type: none"> <li>• HUG - Hôpitaux universitaires de Genève</li> </ul>
<b>Université de Lausanne</b>	<ul style="list-style-type: none"> <li>• Centre hospitalier universitaire vaudois CHUV</li> <li>• Unisanté</li> </ul>
<b>OST – Ostschweizer Fachhochschule</b>	<ul style="list-style-type: none"> <li>• Kompetenzzentrum Demenz</li> </ul>
<b>Kalaidos Fachhochschule</b>	<ul style="list-style-type: none"> <li>• Careum Hochschule Gesundheit</li> </ul>

**Table 4:** List of hospitals/centers affiliated with a swissuniversities member.

After normalization the previously split cells were joined again (“Join multi-valued cells ...”) with the following separators:

- Affiliation: ||
- Affiliation\_copy: ;



In four cases the affiliation to a university (hospital) was added manually after checking the authors.

#### 4. Add column “pages.end” based on “Column 1”:

In almost all records, the eight to last dd-element contains the Source information, e.g. “Krankenpflege: Soins Infirmiers 2020; (9): 78-79. (2p)”. The end page can therefore be extracted with the GREL expression:

```
value.parseHtml().select("dd")[-8].htmlText().rpartition(/\d+\/)[0].rpartition(/\d+\/)[1]
```

Eight values had to be entered by hand after verifying them manually in the full-text of the article.

#### 5. Add column “Accession Number” based on Column 1:

The accession number in the HTM-file corresponds with the value in the XML-element “rec - header - uiTerm”, and is used to merge the information of the two files.

The second to last dd-element contains the Accession Number:

```
value.parseHtml().select("dd")[-2].toString().rpartition(/\d+\/)[1].toNumber()
```

### 4.2.2 Edits on results in XML in OpenRefine

A new project was created by importing the XML file.

#### 1. Adding information from HTM-project

This XML project is enriched with the extracted information from the HTM project, using the GREL function `cell.cross` and the values in columns “Accession Number” and “rec - header - uiTerm”.

Add three new columns based on column “rec - header - uiTerm” in the XML project, using the following GREL expressions:

- **Affiliation normalized:**  
`cell.cross("Krankenpflege HTML", "Accession Number")[0].cells["Affiliation"].value`
- **Affiliation published:**  
`cell.cross("Krankenpflege HTML", "Accession Number")[0].cells["Affiliation_copy"].value`
- **Pages.end:**  
`cell.cross("Krankenpflege HTML", "Accession Number")[0].cells["pages.end"].value`

#### 2. Other edits

- Join multi-valued cells for authors and subjects with separator ||
- Mass edit pattern for ISSN with manual editing after applying text facet to respective column
- Change language according to ISO-code with manual editing after applying text facet to respective column
- Plausibility check for the values in the "rec - header - controlInfo - language" column, as manual spot checks repeatedly revealed discrepancies between the language of the title and the language code (processing steps: 1. add column "language detect" based on column "rec - header - controlInfo - artinfo - tig - at!" with "value.

detectLanguage()", 2. add column "language check" on column "language detect" with "value == cells["rec - header - controllInfo - language"].value", filter this column for values "true", handcode "rec - header - controllInfo - language" after manual check)

- Filter the table in the "affiliation normalized" column for all rows that contain values (total 299 rows) and export result as Excel (.xlsx)
- Filter Excel for all articles with value "Multiple languages" in column "rec - header - controllInfo - language" and splitting the entries in all published language versions of the respective article into separate entries.

#### 4.2.3 Edits with Python script

The following metadata fields were added with a help of a Python Script (Moritz, 2024):

- dc.date.issued
- dc.publisher
- dc.identifier.uri
- publication.status
- dc.rights
- dcterms.type
- file.name

The publisher provides full texts of articles as PDF on their websites:

<https://sbk-asi.ch/de/mitglieder/gemeinsam-stark/fachzeitschrift/>

**Access only via subscription Login:** <https://sbk-asi.ch/Security/login?BackURL=%2Fde%2Fpdf-archiv%2F>

All values in the column "dc.identifiert.uri" are received with web scraping the PDF archive of the publisher. The script searches the PDF archive of the journal using the article title as a search phrase. It then extracts the link from the first search result provided by the PDF archive and adds it to the CSV/Excel file. If no link was found, missing values are coded as "None". No additional plausibility checks on the provided links were made.

## 5 Limitations

Our data is based on data collected by third parties. We trust the data providers' statements that the data is complete. We have not checked whether the data provided fully reflects the publication output of Swiss university members in the two journals in the period mentioned.

## 6 Legal matters

The creators of the reference lists as the basis for data curation are the publisher "Gerontologie CH." or Ebsco (for "Krankenpflege"). In the case that the bibliographic data was created by Swiss publishers, Swiss copyright law applies. As a rule, bibliographic metadata are not considered protected works under the Swiss Federal Copyright Act<sup>10</sup>. Works are defined there as follows: "Works are literary and

<sup>10</sup> [https://www.fedlex.admin.ch/eli/cc/1993/1798\\_1798\\_1798/en](https://www.fedlex.admin.ch/eli/cc/1993/1798_1798_1798/en)

artistic intellectual creations with individual character, irrespective of their value or purpose." It can therefore be assumed that bibliographic metadata are not protected by copyright. However, exceptions are possible. It should be noted that this only applies to the metadata - i.e. the descriptions of the electronic material and not the e-paper itself (full-text as PDF).

In case of bibliographic metadata of the journal "Krankenpflege" the Ebsco licensing agreement<sup>11</sup> applies. We have confirmed with the provider and the publisher that we can use this data in the way described above and publish the CSV/Excel under the Creative Commons licence CC 0 in October 2023.

Regarding the full-texts, authors hold the copyright and transfer certain exploitation rights to the publisher. In the case of articles by authors affiliated with a Swiss university at the time of publication, "Krankenpflege" and "GERONTOLOGIE CH. Praxis + Forschung" grant the respective university the non-exclusive right to deposit a copy of the article in its Open Access repository, in compliance with their self-archiving policy, without necessitating the author's consent.

The self-archiving policies can be found online:

- "GERONTOLOGIE CH. Praxis + Forschung": [https://www.gerontologie.ch/fileadmin/redaktion\\_gerontologie/pdf/Magazin/Zweitveroeffentlichungsrechte\\_GERONTOLOGIE\\_CH\\_D.pdf](https://www.gerontologie.ch/fileadmin/redaktion_gerontologie/pdf/Magazin/Zweitveroeffentlichungsrechte_GERONTOLOGIE_CH_D.pdf)
- "Krankenpflege": [https://sbk-asi.ch/assets/Dokumente-PDF/04\\_Mitglieder/Ihre-Vorteile/2023-Richtlinien-AutorInnen\\_d.pdf](https://sbk-asi.ch/assets/Dokumente-PDF/04_Mitglieder/Ihre-Vorteile/2023-Richtlinien-AutorInnen_d.pdf)

The layout of the PDF edition of "GERONTOLOGIE CH. Praxis + Forschung" consists of double pages that are not separated when split into individual articles, even if other articles are also printed on the pages. It has been clarified with the publisher that repositories may use the full-text files prepared by GOAL for secondary publication in the present form (zip file).

The project uses published personal data of authors (name and affiliation given in the published articles approved by the authors themselves). The data curation and enrichment by the ZHAW normalizes the affiliation on an institutional level and do not include the addition of any further personal information about the authors. Therefore, it does not result in the creation of a personal data profile.

On the contrary, the data provided by GOAL contains less personal information than stated in the original publications (no e-mail address, no indication of the exact organizational unit within the university in case of Gerontologie CH.).

## 7 Credit author statement

Andrea Moritz (Conceptualization, Methodology, Software, Validation, Formal analysis, Data Curation, Writing)

Beatrice Hodel (Methodology, Software, Formal analysis, Data Curation, Writing)

Clemens Trautwein (Writing)

---

<sup>11</sup> <https://www.ebsco.com/license-agreement>

## 8 Acknowledgements

This work was supported by swissuniversities.

We would like to thank our colleague Amira Asfar (ZHAW Zurich University of Applied Sciences) for the English proofreading.

## 9 Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## 10 Attachment code snippet (JavaScript) for Adobe Acrobat

The sample code from the Adobe documentation «JavaScript for Acrobat API Reference» (2021, S. 224, Example 3) has only been adapted in the following points highlighted in yellow:

```

var CopyrightStatus = "True";
var CopyrightNotice = "Copyright(C), Gerontologie CH"
var CopyrightInfoURL = "https://www.gerontologie.ch/impressum"
var meta = this.metadata;
var myXMPData = new XML(meta);
myx = new Namespace("adobe:ns:meta/");
myrdf = new Namespace("http://www.w3.org/1999/02/22-rdf-syntax-ns#");
mypdf = new Namespace("http://ns.adobe.com/pdf/1.3/");
myxap = new Namespace("http://ns.adobe.com/xap/1.0/");
mydc = new Namespace("http://purl.org/dc/elements/1.1/");
myxapRights = new Namespace("http://ns.adobe.com/xap/1.0/rights/");
var p = myXMPData.myrdf::RDF.myrdf::Description;
/*
We test whether this element has a value already, if no, we assign it a value, otherwise
we assign it another value.
*/
if (p.myc::rights.myrdf::Alt.myrdf::li.toString() == "") {
p[0] += <rdf:Description rdf:about=""
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
<dc:rights>
<rdf:Alt>
<rdf:li xml:lang="x-default">
{CopyrightNotice}
</rdf:li>
</rdf:Alt>
</dc:rights>
</rdf:Description>
} else
p.myc::rights.myrdf::Alt.myrdf::li = CopyrightNotice;
/*
Some elements are converted into attributes, so we need to first test whether the
xapRights:Marked attribute is present, if not, we add it in as an element; otherwise, if
the attribute is present, we update the attribute.
Acrobat changes certain elements into attributes; the xapRights:Marked and
xapRights:WebStatement are two such examples, but dc:rights above is one that is not
changed into an attribute.
*/
if (p.@myxapRights::Marked.toString() == "" ) {
p[0] += <rdf:Description rdf:about=""
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:xapRights="http://ns.adobe.com/xap/1.0/rights/">
<xapRights:Marked>{CopyrightStatus}</xapRights:Marked>
<xapRights:WebStatement> {CopyrightInfoURL} </xapRights:WebStatement>
</rdf:Description>
} else {

```

```
p.@myxapRights::Marked = CopyrightStatus;  
p.@myxapRights::WebStatement = CopyrightInfoURL;  
}  
// Convert myXMPData into a string  
myNewXMPStr=myXMPData.toXMLString();  
// and assign it to the document metadata  
this.metadata = myNewXMPStr;
```

## 11 References

- Adobe Inc. (Hrsg.). (2021). JavaScript for Acrobat API Reference. [https://opensource.adobe.com/dc-acrobat-sdk-docs/acrobatsdk/pdfs/acrobatsdk\\_jsapiref.pdf](https://opensource.adobe.com/dc-acrobat-sdk-docs/acrobatsdk/pdfs/acrobatsdk_jsapiref.pdf)
- GOAL - Unlocking the green open access potential. (2024). GOAL. <https://opengoal.ch/>
- Moritz, A. (2024). AnMarlen/goal-data-preprocessing: goal-data-preprocessing package v1.0.0 (v1.0.0). Zenodo. <https://doi.org/10.5281/zenodo.10992164>
- Moritz, A. (2024a). Full-texts of articles in "GERONTOLOGIE CH. Praxis + Forschung" published by Swiss university members [dataset]. <https://doi.org/10.5281/zenodo.10987869>
- Moritz, A., & Hodel, B. (2024). Bibliographic metadata of articles in "Krankenpflege" and "GERONTOLOGIE CH. Praxis + Forschung" published by Swiss university members [dataset]. <https://doi.org/10.5281/zenodo.10987751>
- swissuniversities. (2023, October 18). Mitglieder—Swissuniversities. <https://www.swissuniversities.ch/organisation/mitglieder>
- Trautwein, C., Andres, V., Corredera Nilsson, E., Dobis, D., Flieg, J., Moritz, A., Reymermier, H., Rosenkranz, S., & Simukovic, E. (2022). Dataset underlying the Report of the Publication Analysis GOAL 2020-2021 [dataset]. Zenodo. <https://doi.org/10.5281/zenodo.7063862>