Project Number: **770299**

NewsEye:
A Digital Investigator for Historical Newspapers

Research and Innovation Action
Call H2020-SC-CULT-COOP-2016-2017

D1.9: Data Models (d) (final)

Due date of deliverable: M36 (30 April 2021)

Actual submission date: 10 April 2021

Start date of project: 1 May 2018 **Duration:** 45 months

Partner organization name in charge of deliverable: UIBK-DEA

Project co-funded by the European Commission within Horizon 2020		
Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	-
RE	Restricted to a group specified by the Consortium (including the Commission Services)	-
CO	Confidential, only for members of the Consortium (including the Commission Services)	-

Revision History

Document administrative information		
Project acronym:	NewsEye	
Project number:	770299	
Deliverable number:	D1.9	
Deliverable full title:	Data Models (d) (final)	
Deliverable short title:	Data Models	
Document identifier:	NewsEye-T11-D19-DataModels-d-Submitted-v3.0	
Lead partner short name:	UIBK-DEA	
Report version:	V3.0	
Report preparation date:	10.04.2021	
Dissemination level:	PU	
Nature:	Report	
Lead author:	Günter Mühlberger (UIBK-DEA)	
Co-authors:	Juha Rautiainen (UH-NLF), Axel Jean-Caurant (ULR), Antoine Doucet (ULR), Max Weidemann (UROS), Florian Krull (UIBK-DEA), Günter Hackl (UIBK-IUI)	
Internal reviewers:	Jean-Philippe Moreux (BNF), Martin Gasteiner (UNIVIE)	
Status:		Draft
		Final
	x	Submitted

The NewsEye Consortium partner responsible for this deliverable has addressed all comments received, making changes as necessary. Changes to this document are detailed in the change log table below.

Change Log

Date	Version	Editor	Summary of changes made
02/03/2021	1.0	Günter Mühlberger (UIBK-DEA), Florian Krull (UIBK-DEA), Günter Hackl (UIBK-IUI)	Complete draft version
15/03/2021	2.0	Günter Hackl (UIBK-IUI)	Internal reviews of JP Moreux (BNF) and Martin Gasteiner (UNIVIE) taken into account
09/04/2021	2.1	Günter Hackl (UIBK-IUI)	QM review included
10/04/2021	3.0	Antoine Doucet (ULR)	Minor adjustments and submission

Executive summary

This deliverable presents the cumulative work performed during 3 years in Task T1.1 on 'Data models'. The report outlines the approach with which data and metadata are managed in the NewsEye project. The amount and the variety of data in NewsEye is large: Image data, textual data, layout information, named entities, events and many more data are gathered and created automatically and need to be managed in an efficient way within the project. On the one hand, we follow well-established standards in the field, such as METS/ALTO but also take up on recent developments, such as IIIF (Section 4). This new framework provides a good basis for flexible and future-driven solutions as it is required for a research project with the ambition to set the scene for newspaper digitization in the coming years. However, for some of the questions touched by the NewsEye project also the IIIF framework has no 'standard' solutions available.

Section 2 describes the implementation of the Demonstrator. In Section 3 we present the data models used. We show the data models of the libraries and the data models of Transkribus where Section 3.2.1 explains in more detail how 'articles' are included in the PAGE format. Section 3.3 presents the formats used to store and exchange named entities, named entity links and stances so that the data can directly be used for the training of NLP tools.

Section 4 presents the IIIF data model, while Section 5 introduces key concepts such as 'articles' and also provides a list of structural elements which appear in newspapers and which might play a role when it comes to a more fine-grained annotation of news articles.

Furthermore, we show in Section 6 the structural elements used within NewsEye, which we limit to a feasible subset. Thus, while Section 5 tries to list all possible structural elements in general, this new section restricts this list to a practical but also useful subset of it.

As anticipated for Y2, we planned the evaluation of the suggested data schemes in the project against their actual use with respect to the NewsEye Demonstrator and its implementation. Therefore, in Section 7, we show the connection between the described data models and the Demonstrator.

This deliverable has to be seen in close connection with Deliverable 1.10 on 'Data collection and preservation'. Whereas in this report we lay the focus on the data models in general, deliverable D1.10 describes in more detail the actual process how data are collected within the project. We recommend to read both deliverables side by side in order to get the full picture.

Contents

1 Introduction

As a matter of fact, the digital library community was dominated during the last 15–20 years by the XML based standards set up by the Library of Congress. The most important are:

- MODS: Metadata Object Description Standard
- METS: Metadata Encoding and Transmission Standard
- ALTO: Analyzed Layout and Text Object

These standards are used worldwide, including the participating libraries in NewsEye, which preserve and manage their data by using these standards. However, in the last years a new development was initiated by some well-known libraries, gathered under the hood of the 'International Image Interoperability Framework' (IIIF). The main difference to the conventional XML schemas is the shift in the perspective: Instead of starting with the concept of 'metadata' - as it is natural for analogue libraries – the image itself, or in the notion of IIIF, the 'canvas' is the main focus.

> The 'International Image Interoperability Framework' (IIIF) is a set of shared application programming interface (API) specifications for interoperable functionality in digital image repositories. [...] Using JSON-LD, linked data, and standard W3C web protocols such as Web Annotation, IIIF makes it easy to parse and share digital image data, migrate across technology systems, and provide enhanced image access for scholars and researchers. In short, IIIF enables better, faster and cheaper image delivery. It lets you leverage interoperability and the fabric of the Web to access new possibilities and new users for your image-based resources, while reducing long term maintenance and technological lock in. IIIF gives users a rich set of baseline functionality for viewing, zooming, and assembling the best mix of resources and tools to view, compare, manipulate and work with images on the Web, an experience made portable–shareable, citable, and embeddable.[1]

At the NewsEye kick-off meeting in La Rochelle it became very clear that the NewsEye project - especially in its role as being a vanguard of future digital library application - should go towards this direction. This does of course not mean that METS/ALTO are outdated or no longer usable, but that the main distribution format with which we would like to describe data and meta data within the project would be IIIF and the web annotation framework from W3C. In this way each of the tools, such as layout analysis, text recognition and named entity recognition would provide an additional annotation layer to the source document (body/target).

2 Overall architecture

According to the decisions taken at the kick-off meeting, the work carried out in Task *7.1 Development of a NewsEye Demonstrator* became much more prominent than originally set

[1] https://iiif.io/community/faq/

out in the DoA. Instead of providing 'just' a user interface to the several tools of WP3, WP4 and WP5 it was suggested by ULR to set up the demonstrator on the basis of a digital library / repository application. For reference, the NewsEye demonstrator is briefly described in [1] and in full details in public Deliverable D7.8.

The initial plan was to use the open source library *Samvera* (http://samvera.org), which is based on the well-known FEDORA framework. A good example for *Samvera* is the *Digital Commonwealth* library (https://www.digitalcommonwealth.org).

During the start of the implementation it turned out that using Samvera/Fedora was not a very practical solution. There are several reasons for it:

- Complexity of the repository,
- Slow query processing of Fedora over metadata and documents,
- Unnecessary image repository as we are dealing with IIIF images,
- Too much overhead for a 'proof of concept'.

Thus, the initial plan was changed into a different solution with a web interface based on *Blacklight*[2] (a Ruby plugin). The data is directly indexed with Solr (see Section 7) and the images are fetched directly via IIIF.

This change has an impact on *WP1 Data management* and here especially on *Task 1.1. Data models* and on *Task 1.2. Data collection and preservation*.

Concerning *Task 1.1 Data models*, we rely mainly on IIIF, Solr and the REST APIs of all work packages. With respect to work foreseen in *Task 1.2. Data collection and preservation* the implications of the changed NewsEye Demonstrator implementation is also of eminent importance. The final demonstrator is confined to a fully functional digital library application but without a strong preservation component. Therefore the strategy for the preservation task was changed into uploading all project data to Zenodo[3]. In our opinion, using zenodo is the best way to keep the data alive beyond the end of the project and make it available to other projects as well. The latter is especially very useful for external research groups which can make use of the created NewsEye training data to improve their research results.

[2] https://projectblacklight.org/
[3] https://zenodo.org/

The chart in Figure 1 provides an overview on the main components of the project.

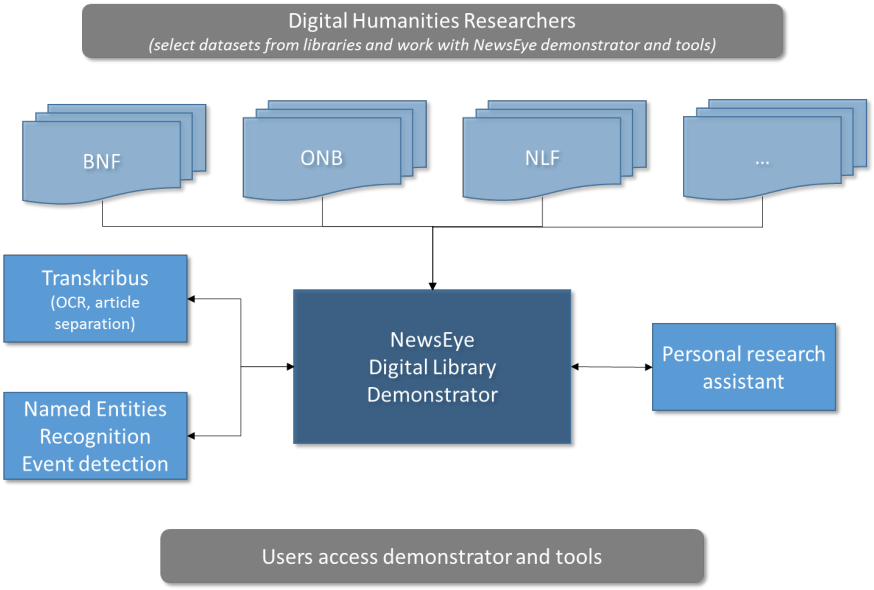


Figure 1: Overview on NewsEye

The technical research within NewsEye has to be seen as a contribution to support DH researchers as well as library users to improve access, research and work with historical newspaper collections. Therefore DH researchers are taking the decision with which data sets they want to work in the project in order to be able to carry out their research. At the first WP6 task meeting in Vienna (07/2018) several proposals were discussed which topics shall be investigated by the DH teams in Finland, Austria and France.

Based on suggestions taken by the DH teams data sets were selected and provided by the participating libraries for ingestion into the NewsEye Digital Library Demonstrator (NDLD). The demonstrator application hosts all data (meta data, images or links to images, annotations) and provides general features for searching and browsing the selected data sets.

During the course of the project the selected data sets were processed with the newly developed or improved tools of the NewsEye project. If – for example – the Automated Text Recognition (ATR) reaches a level of accurateness which is significantly better than the OCR text provided by the libraries within the project, the data sets were processed in Transkribus and the files within the NewsEye Demonstrator updated accordingly.

Besides the NewsEye Digital Library Demonstrator, the Personal Research Assistant is the second main deliverable of the project. It enables users to interactively work with large library collections, to receive information on the documents and collections, to browse through topics but also to get alerted from the assistant on specific issues, topics or research interests.

The Personal Research Assistant is integrated within the Digital Library Demonstrator but will reside outside of NDLD.

3 Data models and interfaces

As we can see from Figure 1 there are several modules where data are rendered and enriched. Clearly defined interfaces are therefore necessary when it comes to exchanging data.We see the following main use cases:

- From libraries to NDLD (and back)
- From NDLD to Transkribus and back
- From NDLD to NER, NEL, stance and event detection and back
- From NDLD to the Personal Research Assistant

Whereas the enrichment of image data with text, layout information, named entities, etc. is a constant process and requires interfaces in both directions, the interfaces between libraries and NDLD as well as between NDLD and the Personal Research Assistant are mainly a one way communication. Of course this does not mean that data enriched in NDLD will not go back to the contributing libraries, but the priority is not to set up a fully-fledged service platform but a demonstrator application showcasing what can be done with cutting edge research.

3.1 From libraries

Digitized newspapers at BNF, ONB and NLF have a similar structure and come with similar formats:

- **General meta data**
  Title of periodical, date of issue, language, number of pages
- **Page images and raw OCR text**
  In all libraries newspapers where OCRed and therefore each image also comes with an OCR file (ALTO)
- **Annotations**
  In some cases also annotations will be made available, e.g. blocks indicating a heading, or articles, or advertisements, etc.

General metadata is used for enabling users to navigate within NDLD according to data sets, newspaper titles, languages and periods.

Page images and raw OCR text from former campaigns such as EU Newspaper are used as a starting point for further analysis.

3.2 From Transkribus and back

The *Transkribus* platform (http://read.transkribus.eu/) is being constantly updated with tools developed by UROS in *WP2 Text recognition and article separation*. Layout analysis and text recognition are already implemented as baseline systems in *Transkribus* to be used as service platform for performing the actual recognition tasks in the NewsEye project.

Data from *Transkribus* have the following characteristics:

- **Layout**
  The pure layout analysis generates blocks, lines and baselines. Blocks contain properties, e.g. if it contains an image, text or a table. Lines of text are polygons including all characters of the line. Baselines are a simplified representation of a line and have turned out to be useful for user interaction since they can be easily created or corrected by human beings (which might be necessary for producing training data)

- **OCR text**
  In contrast to conventional OCR engines such as ABBYY FineReader the line-based processing of ATR engines does not provide coordinates of characters and words. Also properties such as bold, underlined, or the size of characters are not available directly from the ATR engine. However ATR engines provide reliable data on the confidence of each character in the alphabet which can be used to improve processes based on the recognition output, such as named entity recognition. The text therefore comes as Unicode characters on line level and some information about the confidence.

- **Structural annotations**
  Article separation is just one – but probably the most important – layout information of newspapers. Articles – or more generally *news items* – contain a distinct piece of content. Of course we can think on detecting in an automated way many more layout elements, such as headlines, date of publication, place of publication, source of information, etc. but the main focus is on news items. Due to the fact that machine learning methods are applied it is finally the decision of the users which structural annotation is useful for them. Given that enough training data are available the desired output will match with their input.

3.2.1 PAGE format

The PAGE format was introduced in the FP7 Project IMPACT (Improving Access to Text) and is nowadays well established in the computer science and document analysis community. The PAGE format is used within Transkribus to store intermediate processing results as well as performance evaluation aspects to aid the assessment of document image analysis methods on page level. The root structure can be used to point to any kind of data instance which is identified by a namespace and defined by a corresponding XML Schema [2]. The information of an article is stored in the PAGE format as a structure type 'article' to each line with an **'id'** identifying the different articles.

```
<TextLine id="tl_5" primaryLanguage="German"
custom="readingOrder {index:0;} structure {id:a1; type:article;}">
```

Using this ids enables us also to mark articles which span several pages. An example of such a use-case is shown on the next two pictures. Here the second article of the first page (colored in purple) continues on the next page and therefore the same id gets used for all lines belonging to that article no matter which page they are on. The same is true for the 'Feuilleton' (lines are underlined with red color) which starts at the title page and continues on page two.


Neue
Freie Presse.
Morgenblatt.
№ 11085. Wien, Samstag, den 6. Juli 1895.
Wien, 5. Juli.
Feuilleton.
Adolph Wilbrandt: „Die Osterinsel".
Oesterreichischer Reichsrath.

Figure 2: Articles are shown in different colors

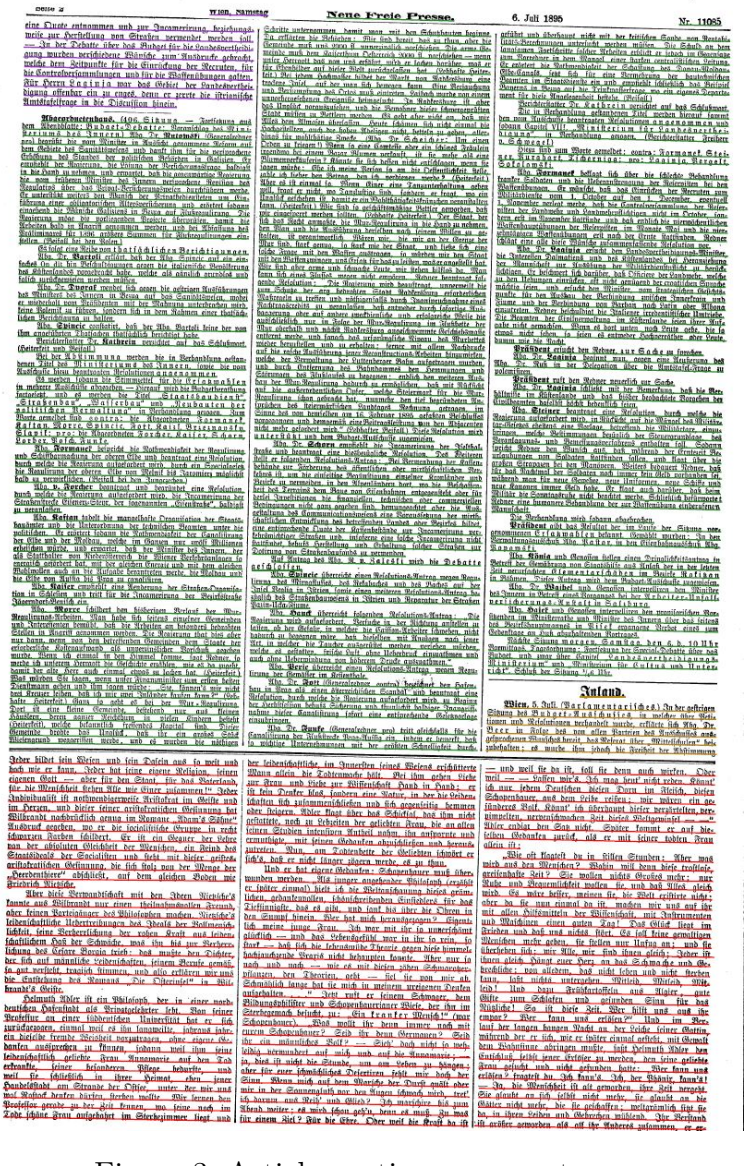


Figure 3: Article continues on next page

This is how we can represent articles which span several pages. Another question is of course whether articles that stretch over several pages can also be recognized automatically. There is no clear answer yet but at least the easy cases should be feasible.

All objects (regions, groups etc.) within the PAGE XML are identified with an ID which has to be unique within the whole XML file. The PAGE XML Schema is further explained based on a Transkribus example below.

3.2.2 Metadata

```
<Metadata>
    <Creator>TRP</Creator>
    <Created>2016-08-30T16:28:46.231+07:00</Created>
    <LastChange>2016-08-31T10:53:22.288+07:00</LastChange>
    <TranskribusMetadata docId="6877" pageId="220277" pageNr="6"
```

```
    tsid="341136" status="NEW" userId="537" imgUrl="https://dbis-thure....;
    fileType=view" xmlUrl="https://dbis-thure...." imageId="206529"/>
</Metadata>
```

3.2.3 Regions

A region reflects a physical object on a page. Regions are defined by their type, outline (polygon), and attributes. Following types are supported:[4] TextRegion, ImageRegion, GraphicRegion, ChartRegion, LineDrawingRegion, SeparatorRegion, TableRegion, MathsRegion, ChemRegion, MusicRegion, AdvertRegion, NoiseRegion, UnknownRegion.Regions can have sub-regions (nested regions), therefore a table-region can include multiple text regions.

```
<TextRegion id="region_1472606927493_80" custom="readingOrder {index:1;}">
    <Coords points="489,103 767,103 767,192 489,192"/>
    <TextLine id="line_1472607237549_87" custom="readingOrder {index:0;}">
        <Coords points="510,131 735,136 734,171 509,166"/>
        <Baseline points="509,161 734,166"/>
        <TextEquiv>
            <Unicode>3. Jan.</Unicode>
        </TextEquiv>
    </TextLine>
    <TextEquiv>
        <Unicode>3. Jan.</Unicode>
    </TextEquiv>
</TextRegion>
```

3.2.4 Reading order

The reading order describes the logical order of text regions. It can have groups and subgroups which can contain either ordered or unordered references to regions.

```
<ReadingOrder>
    <OrderedGroup id="ro_1533643702588" caption="Regions reading order">
        <RegionRefIndexed index="0" regionRef="region_1472606923462_79"/>
        <RegionRefIndexed index="1" regionRef="region_1472606927493_80"/>
        <RegionRefIndexed index="2" regionRef="region_1472606934477_81"/>
        <RegionRefIndexed index="3" regionRef="region_1472606955164_83"/>
        <RegionRefIndexed index="4" regionRef="region_1472606959664_84"/>
        <RegionRefIndexed index="5" regionRef="region_1472606961899_85"/>
        <RegionRefIndexed index="6" regionRef="region_1472606947758_82"/>
    </OrderedGroup>
</ReadingOrder>
```

3.2.5 METS/ALTO and METS/PAGE

In the third year an improvement was done regarding the export of articles and general structural information in Transkribus. This logical structure can be embedded in the METS file and link to the corresponding text lines in the ALTO XML and/or the PAGE XML files. The advantage in comparison to storing the article information directly in the PAGE XML

[4] 2016 PAGE XML Format for Page Content

is that the information is at one place for the whole newspaper issue. The logical structmap is used for this. So in the METS there are not only the physical links to the images and the corresponding text files but also the logical links to the structural information contained in these pages. Since the METS file acts as container for one newspaper issue it would be possible to get e.g. all article headings (if available) very easily for the complete issue and without opening all single PAGE XML or ALTO XML files.

We did a proof of concept for this new export type but did not store all recognised articles in this format since we agreed to use the PAGE XML for that purpose early on. But it will persist in Transkribus as useful export format and every single user of the currently 50,000 registered users can benefit from this NewsEye development.

Here is an example of an exported article:

```
<ns3:structMap TYPE="LOGICAL" LABEL="Logical Structure">
    <ns3:div ID="DIVL1" TYPE="NEWSPAPER">
        ....
        <ns3:div ID="PAGE_1_a5" TYPE="ARTICLE">
            <ns3:div ID="DIVL54" TYPE="TEXTLINE">
                <ns3:fptr>
                    <ns3:area FILEID="PAGEXML_1" BEGIN="tl_57" BETYPE="IDREF"/>
                </ns3:fptr>
            </ns3:div>
            <ns3:div ID="DIVL55" TYPE="TEXTLINE">
                <ns3:fptr>
                    <ns3:area FILEID="PAGEXML_1" BEGIN="tl_58" BETYPE="IDREF"/>
                </ns3:fptr>
            </ns3:div>
            ....
      </ns3:div>
    </ns3:div>
</ns3:structMap>
```

3.3 ConLL and IOB schema for storing NER, NEL, stance training data

A classical problem in information extraction is to recognize and extract mentions of named entities in text. Displaying spans of tokens can be accomplished by the following notation, where each token gets a label that indicates whether a token is a beginning of a named entity, is inside a named entity or is outside a named entity (Inside-Outside-Beginning - IOB format).

3.3.1 IOB scheme

Every token at the beginning of a name span is labeled with a B-prefix. Each token within a name span is labeled with an I-prefix. These prefixes are then followed by a tag indicating the entity type: I-LOC or B-LOC for a location, I-PER or B-PER for a person, I-ORG or B-ORG for an organization and I-HumanProd or B-HumanProd for human production (please have a look to the guidelines of named entities in Deliverable D3.5 for the definition).

Tokens that are not parts of a name span are labeled as O. From this representation, the entity name spans can be recovered unambiguously [3]. In Figure 4 a tagged text phrase in Transkribus is displayed with the according generated file below it. The format for the

internal representation of the named entities is following the data format of the 'CoNLL-2003 Shared Task' [4] of the *Seventh Conference on Computational Natural Language Learning*. We will simply call it CoNLL hereafter. The ConLL format uses the IOB-schema and gets extented with a link to the Wikidata[5] entry, a stance (n = neutral ; + = positive ; - = negative) and EndOfLine|NoSpaceAfter to indicate that the token is at the end of the line and/or there is no space after this token. The latter is true for named entities inside a word - e.g. 'Zürich' in 'Zürichputsch' - or if a named entity is followed by a punctuation mark - e.g. 'Vienna,'. Instead of initially thought we do not use -3 to +3 (from very negative to very positive) as values for stances because the task is difficult enough with just saying that a named entity is mentioned positively or negatively by the author of the article. This is because newspapers are or should be in most cases kept very neutral. And the purpose is not to annotate whether the news is good or bad. For instance, if the article is talking about the sinking of the titanic, the stance with respect to titanic is neutral, even if this is considered bad news. But if the author should criticize the bad design of the titanic then the stance is negative. We need the neutral value because it is not practical to have the stance for every annotated named entity. This would be too much work since the annotator have to read the context and this is very time consuming. So we annotate only a subset of the overall named entity set with stances and therefor we need to distinct between tags with 'neutral' stances and tags with 'no' stances.

If the entity belongs to the entity type person the last column denotes whether the person is the author of the article.

3.3.2 Example in ConLL format

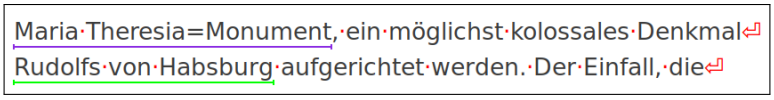


Figure 4: Text tagged in Transkribus

Named entities of Figure 4 exported to ConLL.

```
Maria          B-LOC          https://www.wikidata.org/wiki/Q21234884          n
Theresia=Monument,          I-LOC          https://www.wikidata.org/wiki/Q21234884          n
ein          O
möglichst          O
kolossales          O
Denkmal          O
Rudolfs          B-PER          https://www.wikidata.org/wiki/Q76956          n          author=false
von          I-PER          https://www.wikidata.org/wiki/Q76956          n          author=false
Habsburg          I-PER          https://www.wikidata.org/wiki/Q76956          n          author=false
```

3.4 From NDLD to NER, NEL, stance and event detection and back

The internal representation of named entities, stance and events inside the NDLD is described in Section 3 of Deliverable D3.4. In short, the NER, NEL and stance information is output

[5] https://www.wikidata.org

as ConLL/IOB in the productive process and then converted into a JSON file and read into a SOLR index. This allows the demonstrator to guarantee a fast search. The next two sections show how IIIF can handle this data but it was not used within the project for technical and practical reasons.

3.4.1 Named entity recognition and named entity linking

Named entities can be presented using the IIIF format, as displayed in the example below. In addition to the words representing the entity, we will describe the confidence value of the detection algorithms used. This is also the case for the detected type of entities (person, location, organization, *etc*). When possible, named entities are linked to an external knowledge base (Wikidata[6]), associated with a confidence value. If no entry exist in the knowledge base, one is created to be used as a future reference. Finally, a stance index ranging for instance from -3 to +3 (from very negative to very positive) is associated with an entity mention within a news article.


```
{
   "@context": "http://iiif.io/api/presentation/2/context.json",
   "@id": "http://newseye.eu/68443484/annotation/list/named_entities.json",
   "@type": "sc:AnnotationList",
   "resources": [
       {
           "@id": "http://newseye.eu/68443484/annotation/4809243416439",
           "@type": "oa:Annotation",
           "motivation": "sc:painting",
           "resource": {
               "@type": "cnt:ContentAsText",
               "format": "text/plain",
               "chars": "George Washington"
           },
           "metadata": [
               {
                   "label": "entity_detection_confidence",
                   "value": "0.98"
               },
               {
                   "label": "entity_type_1",
                   "value": "Person"
               },
               {
                   "label": "entity_type_1_confidence",
                   "value": "0.86"
               },
               {
                   "label": "entity_linking_1",
                   "value": "https://www.wikidata.org/wiki/Special:EntityData/Q23.rdf"
               },
               {
                   "label": "entity_linking_1_confidence",
                   "value": "0.96"
```


[6] https://www.wikidata.org

```
            },
            {
                "label": "entity_type_2",
                "value": "Person"
            },
            {
                "label": "entity_type_2_confidence",
                "value": "0.83"
            },
            {
                "label": "entity_linking_2",
                "value": "https://www.wikidata.org/wiki/Special:EntityData/Q2366114.rdf"
            },
            {
                "label": "entity_linking_2_confidence",
                "value": "0.73"
            },
            {
                "label": "entity_stance",
                "value": "0"
            }
        ],
        "on": "http://newseye.eu/68443484/canvas/3320863#xywh=4809,2434,325,39"
    },
    {
        "@id": "http://newseye.eu/68443484/annotation/4809243416440",
        "@type": "oa:Annotation",
        "motivation": "sc:painting",
        "resource": {
            "@type": "cnt:ContentAsText",
            "format": "text/plain",
            "chars": "Washington D.C."
        },
        "metadata": [
            {
                "label": "entity_detection_confidence",
                "value": "0.97"
            },
            {
                "label": "entity_type_1",
                "value": "Location"
            },
            {
                "label": "entity_type_1_confidence",
                "value": "0.74"
            },
            {
                "label": "entity_linking_1",
                "value": "https://www.wikidata.org/wiki/Special:EntityData/Q61.rdf"
            },
            {
                "label": "entity_linking_1_confidence",
                "value": "0.99"
            }
```

```
                {
                    "label": "entity_stance",
                    "value": "-2"
                }
            ],
            "on": "http://newseye.eu/68443484/canvas/3320863#xywh=2568,1458,321,41"
        }
    ]
}
```


3.4.2 Events

The detection of events inside news items is achieved thanks to multilingual approaches based on rhetoric and specificities of the news genre. The main idea is to identify repetition of words in the beginning of news items. For this reason, an event associated to a news item is described by a list of keywords.


```
{
    "@context": "http://iiif.io/api/presentation/2/context.json",
    "@id": "http://newseye.eu/68443484/annotation/list/events.json",
    "@type": "sc:AnnotationList",
    "metadata": [
        {
            "label": "confidence",
            "value": "0.76"
        },
        {
            "label": "keywords",
            "value": "election,representatives"
        }
    ],
    "resources": [
        {
            "@id": "http://newseye.eu/68443644/annotation/4809275489746",
            "@type": "oa:Annotation",
            "motivation": "sc:painting",
            "resource": {
                "@type": "cnt:ContentAsText",
                "format": "text/plain",
                "chars": "election"
            },
            "on": "http://newseye.eu/68443644/canvas/3320863#xywh=4457,2212,478,69"
        },
        {
            "@id": "http://newseye.eu/68443644/annotation/4809275489747",
            "@type": "oa:Annotation",
            "motivation": "sc:painting",
            "resource": {
                "@type": "cnt:ContentAsText",
                "format": "text/plain",
                "chars": "election"
            },
            "on": "http://newseye.eu/68443644/canvas/3320863#xywh=3498,1547,348,65"
```

```
    },
    {
        "@id": "http://newseye.eu/68443644/annotation/4809275489748",
        "@type": "oa:Annotation",
        "motivation": "sc:painting",
        "resource": {
            "@type": "cnt:ContentAsText",
            "format": "text/plain",
            "chars": "election"
        },
        "on": "http://newseye.eu/68443644/canvas/3320863#xywh=5987,3652,647,103"
    },
    {
        "@id": "http://newseye.eu/68443644/annotation/4809275489749",
        "@type": "oa:Annotation",
        "motivation": "sc:painting",
        "resource": {
            "@type": "cnt:ContentAsText",
            "format": "text/plain",
            "chars": "representatives"
        },
        "on": "http://newseye.eu/68443644/canvas/3320863#xywh=5004,2540,645,97"
    },
    {
        "@id": "http://newseye.eu/68443644/annotation/4809275489750",
        "@type": "oa:Annotation",
        "motivation": "sc:painting",
        "resource": {
            "@type": "cnt:ContentAsText",
            "format": "text/plain",
            "chars": "representatives"
        },
        "on": "http://newseye.eu/68443644/canvas/3320863#xywh=3541,1256,531,86"
    },
]
```


3.5 Dynamic text analysis and personal research assistant

Static analysis. In order to perform dynamic text analysis, we first need to perform a static analysis of a large corpus. This prepares the dimensions of analysis for the dynamic analyses that are to be performed on demand. The main example of this in NewsEye is the training of a variety of different types of topic models. This requires textual data for a corpus that is representative of the type of data that will be dynamically analysed, as well as the semantic annotations provided by WP3.

Since a large amount of data must be processed in a bulk, it is not practical to request the entire data set one document at a time using the interfaces provided by the semantic analysis components. Instead, a mirror of the annotated corpora is to be maintained on UH storage for the purposes of model training, and is to be updated when necessary. Semantic annotations are being provided in this static data set in the formats described above.

Dynamic analysis. Dynamic analysis is being performed using a REST API. For example, to make a *document linking* query, the API request specifies a document, or a set of documents, using article IDs common to other components, such as the interfaces in Section 3.4. The JSON result provides a ranked list of related article IDs, with scores and additional metadata to identify the nature of the relation (e.g. particular type of topic model). Other forms of dynamic analysis, such as a *contrastive topic analysis* of two time periods or article collections, uses a similar API interface, also referring directly to articles and annotations using URL-based IDs derived from the relevant components.

Personal research assistant. The personal research assistant consumes both the underlying corpora from Samvera as well as the results of the dynamic text analysis. For this purpose, the access methods provided by the other components of the NDLD are foreseen to be sufficient. Since WP4 requires a UH-local mirror of the corpora, that copy can be accessed to speed up any bulk processing. The personal research assistant then produces textual and graphical results containing annotations referencing the original and annotated corpora. These results are then made available in a structured format that enables them to be embedded into the NDLD user interface.

4 IIIF data model

4.1 IIIF

As already outlined in the introduction of this deliverable we use the IIIF (International Image interoperability Framework) mainly as a distribution format for all data produced within the NewsEye processing workflow. Since IIIF is still in a dynamic stage there are still discussion ongoing how to actually deal with OCR data or how to encode confidences. But we expect that these open issues will be resolved in the coming years.
IIIF allows to combine the presentation semantics and descriptive semantics (metadata) in one format. The IIIF presentation API defines how the structure and layout of a complex image-based object can be made available in a standard manner. The primary requirements for the presentation API are to provide an order for these views, the resources needed to display a representation of the view, and the descriptive information needed to allow the user to understand what is being seen.[7]

4.2 Presentation API

Of special interest is the presentation API of IIIF. It enables users of the NewsEye output, such as libraries but also content aggregators or search engines to build up new services and platforms tailored towards the needs of their specific user groups. A film archive could e.g. integrate historical articles about films and cinemas into their website, or researchers may build up their own archive of source articles connected with their scientific work.

The IIIF manifest represents different types of objects whether it is a newspaper, a book or a painting. Thereby it can be defined that a newspaper has multiple pages.

[7] http://iiif.io/api/presentation/2.1

Canvases represent the views and act as a virtual container for content similar to a PowerPoint slide. In the case of a newspaper, the canvas is a 2d rectangular space with height and width defining the aspect ratios.

The annotations handle all association of content, such as an image related to an article or a transcription can be defined. This information can be used to display it directly in the viewer for information enrichment.

The image API specifies a syntax for web requests to provide images in different sizes, formats and qualities.

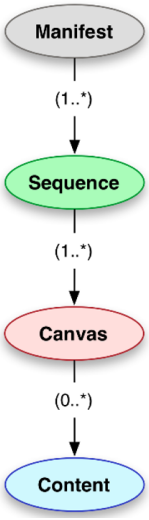


Figure 5: Model of the Presentation API

4.3 IIIF content search API

IIIF specifies the following use-cases searching the annotations within the presentation API[8]:

- Searching OCR generated text to find words or phrases within a book, newspaper or other primarily textual content.
- Searching transcribed content, provided by crowd-sourcing or transformation of scholarly output.
- Searching multiple streams of content, such as the translation or edition, rather than the raw transcription of the content, to jump to the appropriate part of an object.
- Searching on sections of text, such as defined chapters or articles.
- Searching for user provided commentary about the resource, either as a discovery mechanism for the resource or for the discussion.
- Discovering similar sections of text to compare either the content or the object.

[8] https://iiif.io/api/search/1.0/#use-cases

4.4 Machine readable semantic metadata

One way of enhancing the results of the content search API is by adding semantic annotations to IIIF. To create a machine-readable document that semantically describes the resource the IIIF seeAlso property can link to a specific model (MARC record, catalogue metadata, bib.schema.org). For a use-case such as newspapers bib.schema.org would be a good example for a matching semantic annotation, as it combines the core schema.org with a bibliographic extension. The result can be used for search and discovery or inferencing purposes, or just to provide a longer description of the resource.

The manifest links to web pages and other human-readable resources about the thing it represents using the related property. These are links for humans to follow, that a viewer could show in its user interface.

The linked-to resources can point back to the manifest as well. Descriptive metadata at the other end of a seeAlso can link back to the manifest using commonly understood vocabulary [5].

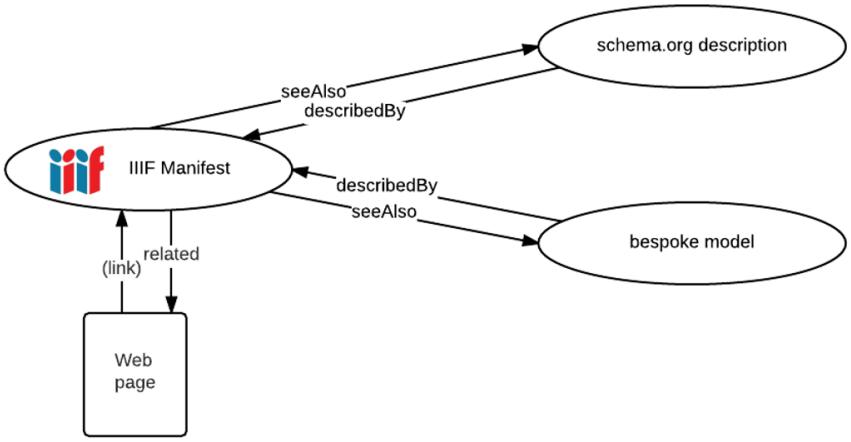


Figure 6: Linking to schema with seeAlso property

4.4.1 Newspaper metadata

Descriptive metadata can be added as well to the collection.

```
{
"@type": "schema:Newspaper",
"editor": "Torsten Kleditzsch ",
"schema:name": "Neue Freie Presse",
"schema:publisher": {
        "@id": "http://www.wikidata.org/entity/Q896460"
}
```

4.4.2 Title section

```
{
"@type": "sc:Range",
  "label": "Neue Freie Presse Morgenblatt",
```

```
  "canvases":
    {
      "@id": "http://anno.onb.ac.at/cgi-content/anno
      ?aid=nfp&datum=19380611&seite=1#xywh=1202,3906,3254,2085"
},
"seeAlso": "https://www.wikidata.org/wiki/Q5707594"
```

4.5 Example of a newspaper article

The following example is based on the work of Tom Crane[9] [5] and enhanced with semantic annotation. Also it illustrates well, how an article starting at the cover of the newspaper and continued on page 3, can be described in JSON. In combination with the semantic annotation it allows search engines to specifically display the content for this article when searched for values defined in the seeAlso property[10].

The annotation allows to define two canvases, which describe the images of the split article. Furthermore the context becomes machine-readable by referring to Wikidata in the 'seeAlso' property. Even though the Wikidata properties are just available as RDF dumps and not as an URI, it will be used in this example for simplicity.

Thereby this article gains a better discoverability by search engines, as 'keywords' or 'backstory' can convey the meaning and purpose of it. In the 'contentLayer' the links refer to the line by line text, where the OCR generated text is contained in the 'chars' property as demonstrated in the example below.

Figure 7: Newspaper article on which Tom Crane's work is based on

```
{
  "@context": "http://iiif.io/api/presentation/2/context.json",
  "@id": "https://tomcrane.github.io/iiif-collector/objects/longer-article.json",
  "@type": "sc:Range",
  "label": "A Profound Warrior For Us All",
  "description": "",
  "canvases": [
    {
      "@id": "https://d.lib.ncsu.edu/collections/canvas/
```

[9] https://tomcrane.github.io/iiif-collector/#objects/longer-article.json

[10] Proposed by Robert Sanderson to use the seeAlso-Property for linking to semantic metadata resources (https://de.slideshare.net/azaroth42/iiif-linked-data)

```
      nubian-message-1995-04-01_0001#xywh=1202,3906,3254,2085",
      "within": {
        "@id": "https://d.lib.ncsu.edu/collections/catalog/
        nubian-message-1995-04-01/manifest.json",
        "@type": "sc:Manifest",
        "label": "Nubian Message, April 1, 1995"
      }
    },
    {
      "@id": "https://d.lib.ncsu.edu/collections/canvas/
      nubian-message-1995-04-01_0003#xywh=313,4540,2388,1579",
      "within": {
        "@id": "https://d.lib.ncsu.edu/collections/catalog/
        nubian-message-1995-04-01/manifest.json",
        "@type": "sc:Manifest"
      }
    }
  ],"contentLayer": {
    "@id": "https://tomcrane.github.io/iiif-collector/
    objects/longer-article-contentlayer",
    "@type": "sc:Layer",
    "label": "Content of `A Profound Warrior For Us All' article",
    "otherContent": [
      "https://tomcrane.github.io/iiif-collector/objects/longer-article-annos1.json",
      "https://tomcrane.github.io/iiif-collector/objects/longer-article-annos2.json"
    ],
    "seeAlso": "https://www.wikidata.org/wiki/Q5707594"
```

4.5.1 OCR text granularity by line

There are several options to annotate images with text. It can be done on page, region, line, word or even character level.[11] In our case we provide text by default on line level. Word level output can be produced however if necessary. This example shows two annotated lines of text in the article. First the article is described as a bib.schema 'Article' and an annotation list, which is a collection of annotations, where each annotation targets the canvas or part thereof.[12] Each line is part of the sequence of objects in the 'resources' property with associated position in the canvas ('on'). Additionally every line can be linked to a schema to further convey context.

```
{
    "@context": "http://iiif.io/api/presentation/2/context.json",
    "@id": "https://tomcrane.github.io/iiif-collector/
    objects/longer-article-annos1.json",
    "@type": "sc:AnnotationList",
    "@textGranularity": "line",
    "within": {
      "@id": "https://tomcrane.github.io/iiif-collector/
```

[11] C.f. Notes from the IIIF working group on Text granularities. https://docs.google.com/document/d/1CCToyJVEr$_G$q2R4GuKV5L51hwpCnC1QOYWcIuf0WU5I/edit#heading = h.mok46p7131n0

[12] https://iiif.io/api/presentation/2.1/#annotation-list

```
    objects/longer-article-contentlayer",
    "@type": "sc:Layer",
    "label": "Content of 'A Profound Warrior For Us All' article",
    "within": {
      "@id": "https://d.lib.ncsu.edu/collections/catalog/
      nubian-message-1995-04-01/manifest.json",
      "@type": "sc:Manifest"
    }
  },
  "seeAlso" : {
          "@context" : "https://bib.schema.org",
      "type" : "Article"
  },
  "resources": [
      {
          "@id": "https://ocr.lib.ncsu.edu/ocr/nu/
          nubian-message-1995-04-01_0001/
          nubian-message-1995-04-01_0001-annotation-list-line/1703,3931,2342,160",
          "@type": "oa:Annotation",
          "motivation": "sc:painting",
          "resource": {
              "@type": "cnt:ContentAsText",
              "format": "text/plain",
              "chars": "A Profound Warrior For"

          },
          "on": "https://d.lib.ncsu.edu/collections/canvas/
          nubian-message-1995-04-01_0001#xywh=1703,3931,2342,160"
          "SeeAlso" : "https://www.wikidata.org/wiki/Q234460"
      },
      {
          "@id": "https://ocr.lib.ncsu.edu/ocr/nu/
          nubian-message-1995-04-01_0001/
          nubian-message-1995-04-01_0001-annotation-list-line/2429,4187,670,158",
          "@type": "oa:Annotation",
          "motivation": "sc:painting",
          "resource": {
              "@type": "cnt:ContentAsText",
              "format": "text/plain",
              "chars": "' Us All"
          },
          "on": "https://d.lib.ncsu.edu/collections/canvas/
          nubian-message-1995-04-01_0001#xywh=2429,4187,670,158"
      }
```


4.5.2 OCR granularity by word

IIIF is in the process of delivering different types of granularity, as the desire to make use of the text in specific way has arisen. The granularity of the text needed, however, varies depending on the use cases as stated by IIIF [13]:

[13] https://iiif.io/community/groups/text-granularity/charter/#introduction

- a word level annotation list for harvesting by Europeana where the aggregator would like to offer word level highlighting for search results
- a line level annotation list for use in Mirador as word level annotation lists can be large for a big newspaper page and reducing the amount of JSON objects can lead to a smoother user experience
- paragraph annotation list for OCR correction where the user wants to have a single box to correct rather than a box per line or word

The following example is based on an article of the 'Neue Freie Presse' 11th of June 1938.[14]


```
{
    "@context": "http://iiif.io/api/presentation/2/context.json",
    "@id": ".....",
    "@type": "sc:AnnotationList",
    "@textGranularity": "word",
    "within": {
      "@id": "...",
      "@type": "sc:Layer",
      "label": "Content of `Wiedergeburt des Theaters' article",
      "within": {
        "@id": "http://anno.onb.ac.at/cgi-content/anno?.../manifest.json",
        "@type": "sc:Manifest"
      }
    },
    "resources": [
        {
            "on": "http://anno.onb.ac.at/cgi-content/anno?aid=nfp&datum=19380611...",
            "@type": "oa:Annotation",
            "motivation": "sc:painting",
            "resource": {
                "@type": "cnt:ContentAsText",
                "format": "text/plain",
                "chars": "Wien"
            },
            "on": "http://anno.onb.ac.at/cgi-content/anno?aid=nfp&datum=19380611...",
            "seeAlso" : "https://www.wikidata.org/wiki/Q1741"
        },
        {
            "on": "http://anno.onb.ac.at/cgi-content/anno?aid=nfp&datum=19380611...",
            "@type": "oa:Annotation",
            "motivation": "sc:painting",
            "resource": {
                "@type": "cnt:ContentAsText",
                "format": "text/plain",
                "chars": ","
            },
            "on": "http://anno.onb.ac.at/cgi-content/anno?aid=nfp&datum=19380611...",
            "seeAlso" : "https://www.wikidata.org/wiki/Q161736"
        },
```


[14] http://anno.onb.ac.at/cgi-content/anno?aid=nfp&datum=19380611&seite=1#xywh=1202,3906,3254,2085

```
{
    "on": "http://anno.onb.ac.at/cgi-content/anno?aid=nfp&datum=19380611...",
    "@type": "oa:Annotation",
    "motivation": "sc:painting",
    "resource": {
        "@type": "cnt:ContentAsText",
        "format": "text/plain",
        "chars": "11."
    },
   "on": "http://anno.onb.ac.at/cgi-content/anno?aid=nfp&datum=19380611...",
    "seeAlso" : "https://www.wikidata.org/wiki/Q205892"
},
{
    "on": "http://anno.onb.ac.at/cgi-content/anno?aid=nfp&datum=19380611...",
    "@type": "oa:Annotation",
    "motivation": "sc:painting",
    "resource": {
        "@type": "cnt:ContentAsText",
        "format": "text/plain",
        "chars": "Juni"
    },
    "on": "http://anno.onb.ac.at/cgi-content/anno?aid=nfp&datum=19380611...",
    "seeAlso" : "https://www.wikidata.org/wiki/Q205892"
}
```


Due to the fact that the IIIF Presentation API does not directly address semantic tagging in its specification, there are many proposed ways to achieve this goal. For example Gene Loh is using a region of a painting tagged with the 'Dublin Core Metadata Initiative' (DCMI) dcterms:subject associative relationship (rel—predicate) to a term in the 'Getty AAT vocabulary' (href—object) [6].


```
{
  "@context": "http://iiif.io/api/presentation/2/context.json",
  "@id": "https://demo.linkedcanvas.org/annotatedImage1/annotation/anno1",
  "@type": "oa:Annotation",
  "motivation": "sc:painting",
  "resource": {
    "@id": "https://demo.linkedcanvas.org/annotatedImage1/tag1",
    "@type": "oa:SemanticTag",
    "rel": "dcterms:subject",
    "href": "http://vocab.getty.edu/aat/300055165"
  },
  "on": "https://demo.linkedcanvas.org/annotatedImage1/canvas/p1#xywh=100,100,500,300"
}
```


4.6 Confidence values

Since machine learning methods play a dominant role nowadays we are dealing not only with the output of data, such as text, layout or named entities, but the output is usually also connected with a confidence value. Such a confidence value can be a valuable input for other tools since it is much richer than the output finally provided to the user. But since this is mainly useful for handwritten documents with much higher error rates we can omit

to store these confidence values. This avoids overhead for calculating the values as well as saves storage and reduces the complexity also for the Demonstrator. So we only show how the confidence values could be stored in such a case. From the point of view of IIIF such an output is an additional annotation with very specific properties.

These confidence values are stored in a so called confidence matrix (ConfMat), the output format of the automated text recognition (ATR). Per row a ConfMat describes a probability distribution over a set of classes present in a model. These classes are represented by the characters occurring in the training data set on which the ATR model is based on. Figure 8a shows an example of a ConfMat where the first column belongs to a special garbage channel, called Not-a-Character (NaC), and the second column to the space.

ConfMats are already being used successfully for keyword spotting (KWS) and language modeling tasks as described in [7]. Again this is valuable much more for handwritten than for printed documents. If the error rate is in the range of 1% the user can find the words with a normal SOLR search as well and does not need KWS.

To use the ConfMat for more advanced tasks like named entity recognition (NER) or topic modeling we need to export it in a suitable format. This is achieved by saving them in CSV files following the RFC-4180 specification using the UTF-8 encoding scheme where the filenames are given by "line_id.csv". Thus, a clear assignment between ConfMat and textlines is ensured. Furthermore the files are stored in US-en format and the confidences are given in exponential notation with a significant length of 6. In Figure 8b a snippet of such a CSV file is given for the example in Figure 8a. Another suitable format to store the confidence values is ALTO XML. The Schema Version 4.0[15] introduced the character based text description with new 'Glyph' element and its subelement 'Variant' to be able to handle these kind of information stemming from new machine learning methods. Here is a part of the documentation from the schema definition for this new addition:

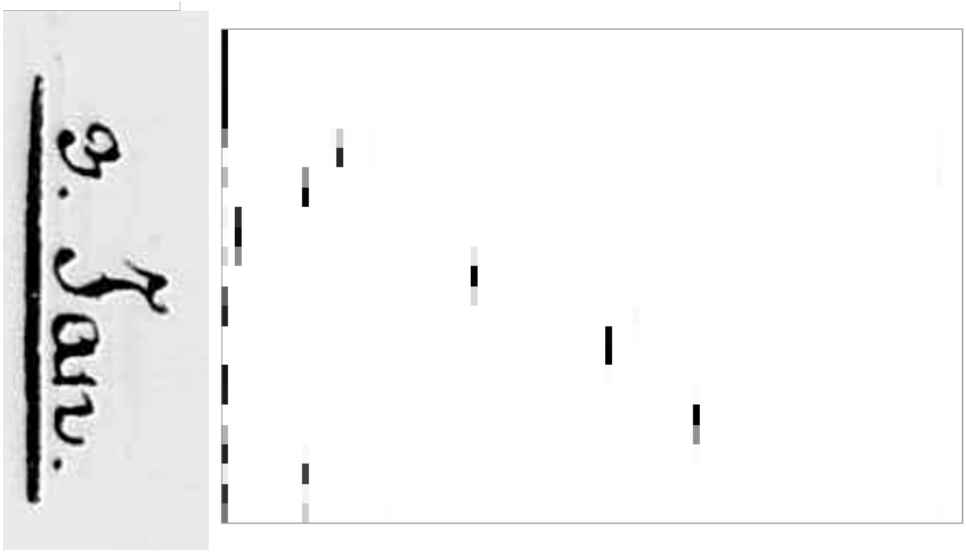
> *In order to reproduce the decision of the OCR software, optional characters must be recorded. These are called variants. The OCR software evaluates each variant and picks the one with the highest confidence score as the glyph. The confidence score expresses how confident the OCR software is that a single glyph had been recognized correctly.*

5 Main concepts within newspapers

This section deals with the main concepts necessary to describe the full richness of newspaper content. Though newspapers developed over 300 years we believe that with a handful of concepts and rules it is possible to find a path through this complex resource. These concepts are:

- Newspaper sections
- News items (or simply 'articles')
- Structural elements

[15] https://www.loc.gov/standards/alto/v4/alto-4-0.xsd



(a) Line of handwritten text: "3. Jan." used as input to the ATR model and the corresponding output of the Neural Network. Black means high confidence, white means very unlikely. Rows correspond to positions in the image, columns correspond to classes (i.e. characters).

```
1  "","»   "," ","!","""",%,&,',(,),",",-,.,/,0,1,2,3,4,5,6,7,8,9,:,;,=,?,A,B,C,
2  9.96193E-1,1.0302E-5,1.36204E-8,1.79394E-5,1.49948E-5,1.25145E-5,1.31007E-6,1
3  9.95433E-1,8.00895E-6,4.01826E-9,1.33529E-5,1.06212E-5,8.87107E-6,7.53473E-7,
4  9.95433E-1,5.76895E-6,1.98401E-7,8.00895E-6,5.791E-6,6.01618E-6,3.41817E-7,5.
5  9.97714E-1,5.05948E-6,5.64702E-6,4.98286E-6,3.66952E-6,5.03637E-6,1.22499E-7,
6  9.97714E-1,5.49403E-6,9.19495E-6,4.50892E-6,3.87378E-6,8.56513E-6,6.64348E-7,
7  6.35595E-1,5.28436E-6,6.9229E-7,4.07695E-6,3.86492E-6,1.51674E-5,1.00305E-4,1
8  2.61514E-2,2.65938E-7,1.1551E-8,2.54039E-7,1.83267E-7,1.41285E-6,1.34552E-4,8
9  3.89752E-1,4.44742E-6,5.86659E-8,7.59824E-6,4.81102E-6,2.05959E-5,1.22313E-4,
10 4.04289E-3,3.22068E-8,1.53303E-6,7.28042E-8,3.13821E-8,7.73273E-8,1.43348E-7,
11 7.30273E-2,2.25533E-7,9.19501E-1,2.89883E-7,1.55897E-7,2.92773E-7,4.18088E-8,
12 1.225E-2,3.96346E-8,9.87867E-1,3.70892E-8,2.51723E-8,3.68916E-8,1.50866E-9,4.
13 2.55012E-1,1.21568E-6,5.9523E-1,9.99992E-7,8.4483E-7,1.28924E-6,1.632E-7,2.29
14 3.45488E-5,1.56732E-10,1.46443E-6,1.33835E-10,1.07762E-10,3.43384E-10,2.16595
15 7.76828E-1,4.03671E-6,1.26489E-4,2.24846E-6,2.12503E-6,1.14895E-5,1.28433E-6,
16 9.54526E-1,3.02076E-6,2.66142E-6,1.16573E-6,1.16218E-6,7.35301E-6,1.60606E-6,
```

(b) A snipped of the corresponding CSV file, where the first row represents the channels. The other rows depict the confidences of the channels for a given timestep in the image.

Figure 8: Example of a ConfMat and the corresponding export CSV format.

In the following we will describe in more detail the concepts and also provide examples for a better understanding.

5.1 Sections within newspaper issues

A newspaper issue may consist of several sections which include several other sections and news items. It is not easily possible to draw a clear line between the hierarchies in all cases.

The main criteria for a section is that it is repeated over several issues of a newspaper. In some way it is a 'placeholder' filled with the articles and other content gathered from the previous edition of the newspaper to the actual one. Typical newspaper sections are 'Foreign affairs', 'Job announcements' or 'List of persons born in the last week'. Newspaper sections where introduced early. Their origin is the wish of editors to order news items by place and date of their origin.

Since sections are structures which overlap not only pages but also complete issues they are currently excluded from the layout analysis and ground truthing tasks.

5.2 News items

The main concept within newspapers is 'articles' or as we would like to call them 'news items'.

A news item is a distinct piece of content within a newspaper which can be clearly separated from other news items by its content. In order to understand a news item no further context or information is necessary.

The main reason to use the term 'news item' is that we need a concept which also covers job announcements, commercial advertisements, or letters to the editor in the same way as what is usually understood as an article, e.g. a report about a car accident, a crime case, or a commentary on the current political situation.

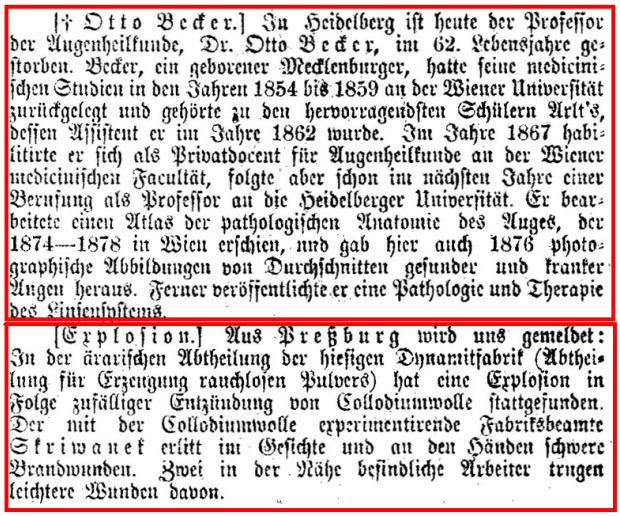
The main criteria to make a distinction between two news items is their contents. A paragraph in an article, or a row in a stock exchange table are pieces of a news item, but are not pieces on their own. To be fully understood they need some context which is provided by the 'rest' of the item.

Figure 9 shows two news items separated from each other with just minimal layout effort. The heading is put into square brackets, indicating that the news editor was fully aware of the fact that the heading is not directly part of the story, but on a 'meta-level'. The letter spacing was used for visually emphasizing the start of the news item.

5.3 Structural elements

Structural elements are the third main concept of our scheme. They are necessary to 'materialize' the content of newspapers in a way that the user can understand its meaning on a meta-level, say without 'close' reading. Examples of structural elements are headlines, sub-headlines, captions, copyright notes, paragraphs, tables, and many more such elements.

Newspaper structural elements are defined by their functionality for structuring the content of a newspaper issue. E.g. headlines raise the attention of a reader and inform him or



[† Otto Becker.] In Heidelberg ist heute der Professor der Augenheilkunde, Dr. Otto Becker, im 62. Lebensjahre gestorben. Becker, ein geborener Mecklenburger, hatte seine medicinischen Studien in den Jahren 1854 bis 1859 an der Wiener Universität zurückgelegt und gehörte zu den hervorragendsten Schülern Arlt's, dessen Assistent er im Jahre 1862 wurde. Im Jahre 1867 habilitirte er sich als Privatdocent für Augenheilkunde an der Wiener medicinischen Facultät, folgte aber schon im nächsten Jahre einer Berufung als Professor an die Heidelberger Universität. Er bearbeitete einen Atlas der pathologischen Anatomie des Auges, der 1874—1878 in Wien erschien, und gab hier auch 1876 photographische Abbildungen von Durchschnitten gesunder und kranker Augen heraus. Ferner veröffentlichte er eine Pathologie und Therapie des Linsensystems.

[Explosion.] Aus Preßburg wird uns gemeldet: In der ärarischen Abtheilung der hiesigen Dynamitfabrik (Abtheilung für Erzeugung rauchlosen Pulvers) hat eine Explosion in Folge zufälliger Entzündung von Collodiumwolle stattgefunden. Der mit der Collodiumwolle experimentirende Fabriksbeamte Skriwanek erlitt im Gesichte und an den Händen schwere Brandwunden. Zwei in der Nähe befindliche Arbeiter trugen leichtere Wunden davon.

Figure 9: Two typical news items (Neue Freie Presse, 8. February, 1890, p5)

her about the main content of a news article. The copyright note ('by-line') at the beginning of a news item provides the information, who, where and when an article was written, the caption explains the content of a picture, table or chart.

The main role of structural elements is to provide a kind of 'meta-information' to the reader. Structural elements aim at supporting the reader in understanding the content intuitively and being able to navigate through the complex content of a newspaper. It is therefore very interesting to see how structural elements develop over time in order to cope with the ever growing quantity of content published in a newspaper but also with the increasing complexity of the newspaper format itself.

Due to the fact that the repertoire of structural elements was developed over a long period of time a specific 'message' is associated with many structural elements. E.g. even if we look at a newspaper from far away, or in a completely foreign language, we can understand the semantics of elements, such as headlines, sub-titles, caption lines, etc. even when we are not able to read and understand a single word. It is exactly this aspect that makes us so confident that this 'code' can be deciphered with machine learning algorithms if large amounts of training data are available.

5.4 List of structural elements

This list is a first attempt to cover the most important layout elements of (historical) newspapers. The list provides for each element a definition, the functional value, a note on the chance for automated capturing and examples taken mainly from Austrian newspapers. In addition some 'schema.org'[16] properties with corresponding definitions are proposed which might be used later in the project.

In order to make the list as straightforward as possible, we have left out all those structural

[16] *Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.* This description was taken from https://schema.org/: 'Welcome to Schema.org'

elements which contain information which is recorded anyway during the digitisation process and which only play a minor role in the document analysis of newspapers. E.g. the newspaper title itself, the responsibility statement, the date, issue number or page number need not to be mentioned here explicitly.

5.4.1 Title section

Definition
The first part of an issue containing metadata such as the name of the newspaper, date of publication, imprint information, issue number, etc. Usually this type of information is dependent on the legal obligations which need to be fulfilled by a newspaper publisher.

Functional value
The title section can be regarded as a 'shrunk' title page of a book. Usually the descriptive meta data contained in the title section, such as name of the newspaper, edition, publisher, issue date, etc. are captured as part of a library catalogue or in beforehand during the scanning process. From the point of view of document analysis the title section needs to be captured as a whole but not in detail.

Automated capturing
The title section can easily be detected automatically since its location is highly regular and the text itself is repeated in every issue with only slight differences.

Example
Figure 10 shows a classical title section from the 'Neue Freie Presse', 11th June 1938.

Figure 10: Title section, 1938

5.4.2 Running title

Synonyms
Header, column title

Definition
The running title appears at the top of a newspaper page and typically includes the (short)

title of the newspaper, the page number, the issue number, a heading and the date of publication.

Value
The functional value of the running title is to provide the user with a quick orientation on which newspaper issue s/he is actually reading (number, date) and maybe some information on the content of the page (e.g. if a section heading is included). The benefit of running titles for document analysis is rather low since nearly all the information is already available and needs not to be marked on every page of a newspaper. As already mentioned the only exception can be seen if running titles contain also a subject heading indicating the content of the page, e.g. 'Sports'. Then this information can be extracted and be used to make a general classification of the content of this page.

Automated capturing
Running titles can easily be located and extracted automatically, since they have a clearly defined position within the overall layout of a newspaper page and their content can also be predicted in a highly regular way.

Examples
The 'Wiener Zeitung' (first published issue in 1703) introduced its first running title in 1784. It contained for more than 100 years just the page number, from 1876 onwards, also the issue number, title, and date are displayed.

5.4.3 Heading

Synonyms
Head, main title, title

Definition
News items of the category 'information' and 'opinion' usually come with a title indicating briefly the content of the article. Large articles may have several titles, such as top heading, a heading, a sub-heading and several inside-headings. Other news items falling under the advertisement or entertainment category often do not follow this regime and have no clear 'heading', but are indeed 'section headings', e.g. if we think on 'job offers' or 'family news'.

Value
The functional value of headings is twofold: First to mark the beginning of a news item by providing an 'eye catcher' which is in the easiest case a word in bold, or in brackets, or in later times a word with a large font size. Second to attract the attention of a reader with a short message to actually select this article (instead of other, less important ones).

Automated capturing
In modern newspapers the capturing of headings is often based on the different font size compared to the running text. With historical newspapers the situation becomes much harder since the font size is often just a few points larger and this might not be sufficient

for automated detection. Nevertheless as we will point out below there are several ways to detect the headline in an indirect way, by taking benefit from some background knowledge. E.g. repeated section headings, the mentioning of the place and date of the news and the copyright statement may be utilized for this purpose. Machine learning approaches which are able to take benefit from such context information in an automated way will likely cope with this situation.

Schema
Headline of the article.
https://schema.org/headline

Examples
It took until the middle of the 19th century for newspapers in Central Europe to indicate the start of a news item with a specific heading. Before that time it was the place of origin (see Figure 11: coverage note spatial) and the date of the news which was used. 'Aus der Schweiz' indicates the section, 'Genf 26. May. [1750]' shows from where and when the news comes. It is important to understand that 'Genf 26. May' is not so much a heading, but contains metadata about the news itself.

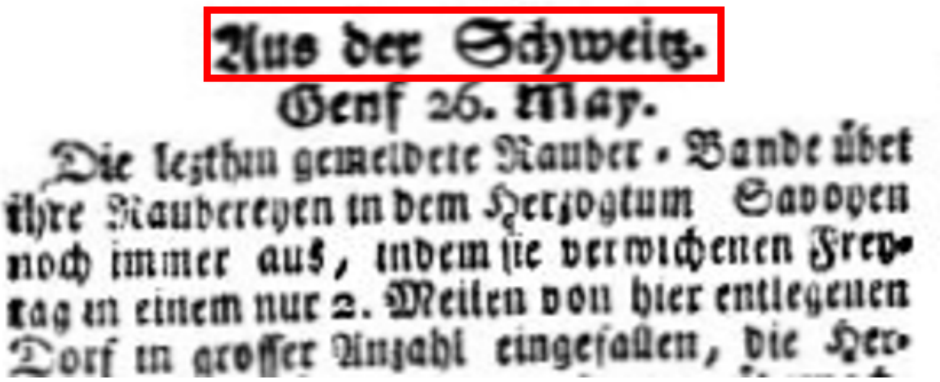

Aus der Schweitz.
Genf 26. May.
Die lezthin gemeldete Rauber-Bande übet ihre Raubereyen in dem Herzogtum Savoyen noch immer aus, indem sie verwichenen Freytag in einem nur 2. Meilen von hier entlegenen Dorf in grosser Anzahl eingefallen, die Her-

Figure 11: Section heading and coverage note spatial from 1750

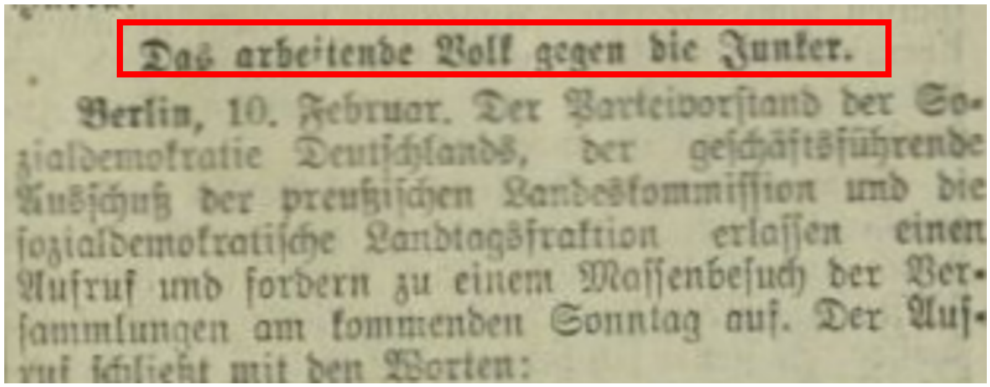
100 years later we still have a very similar situation (see Figure 12): A section heading 'Inland' indicates the start of this section, 'Wien' is the place of origin, the 12th June the date, but now a (short) first sentence summarizes the content of this article: 'Der Bruch im czechischen Lager' (The crash in the Czech party). This first sentence is set in spaced letters and put into brackets in order to separate it from the actual running text which is a clear indication that it is the heading of this news item.

Inland.

Wien, 12. Juni. (Der Bruch im czechiſchen Lager.) Wenn die Jungczechen vermeinten, durch die projectirte Aufſtellung einer eigenen Candidatenliſte auf ihre altczechiſchen Gegner einen heilſamen Druck auszuüben und ſie einer Verſöhnung zugänglicher zu machen, ſo können ſie heute ſchon ſehen, daß ſie ſich hierin gewaltig irrten. Die

Figure 12: Similar situation: Die Presse, 1874

It takes until the 20th century that the modern form is found with a large printed heading indicating the start of the news and a coverage note with additional information. Figure 13 shows the heading 'Das arbeitende Volk gegen die Junker' and 'Berlin, 10. Februar' as additional information.



Das arbeitende Volk gegen die Junker.

Berlin, 10. Februar. Der Parteivorſtand der Sozialdemokratie Deutſchlands, der geſchäftsführende Ausſchuß der preußiſchen Landeskommiſſion und die ſozialdemokratiſche Landtagsfraktion erlaſſen einen Aufruf und fordern zu einem Maſſenbeſuch der Verſammlungen am kommenden Sonntag auf. Der Aufruf ſchließt mit den Worten:

Figure 13: Classical start of a news item, Arbeiterwille, 1910

5.4.4 Sub-heading

Definition
A title that follows the main heading of an article and which provides some additional information.

Functional value
The value is high since the content of the news is often explained in more detail.

Automated capturing
Important news items appear most frequently with a sub-heading which comes in a specific layout.

Schema
A secondary title of the CreativeWork.
https://schema.org/alternativeHeadline

Examples
Sub-headings appear rather late in the first half of the 20th century. An example from the 'Deutsche Zeitung' in Temesvar (Romania) from 12th March 1938 is shown in Figure 14.


Preis 3 Lei

Banat=er

Deutsche Zeitung

20. Jahrgang. Timişoara, Samstag, 12. März 1938 Nr. 57

Ganz Österreich im Fieber

Die Volksbefragung in Abwesenheit von Seyß-Inquart und Gleise-Horstenau beschlossen

Der deutsche Standpunkt: Der Beschluß ist verfassungswidrig und steht im Widerspruch zu dem Abkommen von Berchtesgaden

Wien, 11. März

Einer Meldung aus Graz zufolge, hat dort die Leitung der Nationalsozialisten die Mitglieder aufgefordert, sich der Stimmabgabe bei dem Volksentscheid zu enthalten. Auch die volkspolitischen Referenten in Graz und Linz, die sich um die Aussöhnungsaktion bemühten, haben ihren Bedenken hinsichtlich mehrerer Fragen im Zusammenhang mit der Art und Weise des Aufrufes zur Volksabstimmung Ausdruck gegeben. Die Nationalsozialisten erklären offen, daß Schuschnigg sich gegen die Verfassung vergangen hat, da ein so wichtiger Staatsakt nicht über

die Behauptung sei nicht stichhältig, wonach sie dem Abkommen mit Deutschland vom Juli 1936 und Februar 1938 widerspreche. Im Gegenteil, der Bundeskanzler wolle, das österreichische Volk möge seine Zustimmung zu der Fortsetzung der klaren politischen Richtlinien Schuschniggs geben.

Man rechnet in Wien damit, daß bei der Volksabstimmung in Oesterreich etwa 4.2 Millionen Stimmen abgegeben werden. Die österreichische Regierung weist die Beschuldigung der Nationalsozialisten, wonach die Volksabstimmung gegen die Bestimmungen

Unter 60 Prozent — Schuschniggs Sturz

Wien, 11. März.

Wie aus politischen Kreisen verlautet, wird Schuschnigg in dem Fall abdanken, wenn bei der Volksabstimmung die Ja-Stimmen weniger als 60 Prozent ausmachen. In diesem Fall sei es aber zweifellos, daß dann eine ganz nationalsozialistische oder eine Regierung mit überwiegend nationalsozialistischen Mitgliedern an die Spitze des Staates gelangen werde. Die Folge davon werde die frühere oder spätere Verwirklichung des Anschlusses sein.

Figure 14: Top heading, heading and sub-heading for an extraordinary important news (annexation of Austria)

5.4.5 Inside-heading

Synonyms
Sub-heading

Definition
Larger news items are sometimes structured with headings directly within the text body. In contrast to highlighted words or phrases which do not disrupt the running text actual inside headings are distinct pieces of text and are therefore not part of the actual running text. The main function in this case is therefore to highlight the importance which is in other cases done with spaced letters, bold or italic. So, the decisive criteria to distinguish between pure highlighted words and actual inside headings is: If they do not interrupt the normal reading order we speak just of highlighting (expressed with a larger font size, and text styles), if this piece interrupts the flow of text then it is an inside-heading.

Functional value
Since inside-headings are embedded particles within the text they 'destroy' the running text. This may have some negative affect if the text is parsed, e.g. with parsers structuring text into sentences and applying grammatical categories (POS tagging) to the words.

Automated capturing
Since inside-headings are not used in general but mainly by specific newspapers for a given period of time it may be possible to detect and extract them in an automated way.

Schema
A secondary title of the CreativeWork.
https://schema.org/alternativeHeadline

Examples
Figure 15 shows a good example of a pure highlighting which may look like an inside-heading.

Schülermordtragödie.

(Eigenbericht.)

Berlin, 10. Februar.

Unter unvermindertem Andrange des Publikums begann heute früh der zweite Tag der Verhandlung über die Steglitzer Schülertragödie. Rechtsanwalt Dr. Frey beantragte die Vorladung des Polizeipräsidenten Dr. Weiß als Zeugen und Sachverständigen. Dr. Weiß habe bei den Vernehmungen über das Verhalten der Zeugin Hilde Scheller außergewöhnliche Wahrnehmungen gemacht. Lächelnd, freudig und strahlend habe sie das, was geschehen sein sollte, mit einer Sicherheit, wie sie Dr. Weiß noch niemals vorgekommen sei, teilweise in Abrede gestellt. Noch nie habe Dr. Weiß, wie er sich selbst ausdrückte, einen derart erschütternden Eindruck von der Unglaubwürdigkeit eines Menschen erhalten. Nach kurzer Beratung beschloß das Schwurgericht, die Entscheidung über diesen Antrag zurückzustellen, bis die Vernehmung der Zeugin Hilde Scheller abgeschlossen ist.

Dann wurde

die Beweisaufnahme

fortgesetzt und Hilde Scheller weiter vernommen. Sie schildert im einzelnen, wie sich das Zusammentreffen aller Beteiligten am Abend vor der Tat abgespielt habe. Stephan habe wohl ihren Bruder gesehen, aber dieser nicht ihn. Sie habe dann Stephan heimlich mit in die Wohnung genommen und sei mit ihm ungefähr dreiviertel Stunden in der Kammer ihres Bruders gewesen. Dort hätten sie sich unterhalten und auch geküßt.

Erster Staatsanwalt Steinböck: Ich möchte anheimstellen, den Ausschluß der Öffentlichkeit zu beschließen. Es ist wohl zu erwarten, daß der Verteidiger über diese Angelegenheit noch weitere Fragen stellen wird, und ich möchte nicht, daß die sexuellen Dinge wieder in voller Öffentlichkeit erörtert werden.

Verteidiger Rechtsanwalt Dr. Frey: Ich

den und die Küche sei voll Zigarettenqualm gewesen. Günter habe eine Pistole in der Hand gehabt. Ich habe gesagt: „Wo habt ihr denn hingeschossen?" Da sagten beide: „In die Tür." Dann sagte Paul, er wäre beinahe erschossen worden.

Vorsitzender: Das war doch ungewöhnlich?

Zeugin: Nun, sie hatten ja schon öfter geschossen.

Die Zeugin gibt an, daß neben den Gläsern und Flaschen auch Briefe herumgelegen seien. Sie sei aber so müde gewesen, daß sie sich nicht darum gekümmert hätte.

Im Schlafzimmer fragte Stephan, was los sei. Ich sagte: „Gar nichts, sie haben bloß geschossen." Darauf meinte Stephan: „Nichts weiter als geschossen, da dank ich sehr." „Ach," sagte ich, „sie spielen immer mit Pistolen." In dieser Art haben wir uns unterhalten. Als er fragte, wer denn draußen mit ihnen sei, habe ich gesagt: „Den kennst du nicht."

Vorsitzender: Warum haben Sie nicht die Wahrheit gesagt?

Zeugin: Weil Günter mit Hans verfeindet war. Günter hat mit einem fremden Herrn verschiedene Ausflüge gemacht und Hans hat das seiner Mutter erzählt. Darauf ist Frau Stephan zu Günters Mutter gekommen und hat das wiedererzählt. Deshalb ist Günter mit Stephan böse gewesen. Die Zeugin bestreitet dann, daß sie Paul Krantz Treue geschworen habe und meint, Hans hätte überhaupt nicht so lange bei ihr bleiben sollen, sie habe ihn beim Fenster herauslassen wollen. Es sei aber schon zu hell gewesen. Es hätte einen komischen Eindruck gemacht, wenn jemand aus dem Fenster gestiegen wäre. Der Vorsitzende fragt sie dann, ob sie sich erinnert, daß sie dem Untersuchungsrichter gesagt habe, sie habe Hans wirklich so gern bekommen, daß ihr alles egal war. Die Zeugin erzählt dann weiter, daß am Morgen, als sie einmal herauskam, Paul sie festhalten wollte und sie bat, einen Augenblick bei Paul und Günter zu bleiben. Sie

Figure 15: Arbeiterwille, 11th February 1928

5.4.6 Top heading

Synonyms
Top title, roof title

Definition
Similar to the sub-title of an article and providing the same structural functionality there might be a title above the main title. An example of such a top heading is contained in Figure 14.

Value
The value is very similar to the sub-heading since the content is explained in more detail.

Automated Capturing
Top titles were introduced relatively late in newspapers. They always appear above the main title and are therefore rather easy to detect automatically.

Schema
A secondary title of the CreativeWork.
https://schema.org/alternativeHeadline

5.4.7 Lead paragraph

Synonyms
Intro, Introduction

Definition
Usually the first paragraph(s) of a (larger) article providing an overview of the content of the article.

Functional value
The lead is a kind of abstract of an article. It appears only within larger news articles and may therefore be used for displaying purposes. Since the lead does not break up the running text, but is part of it, its detection is less important compared to headings and other structural elements. A lead must not be mixed up with real 'summaries' as they appear separately as preview of the content of a newspaper issue.

Automated capturing
The lead is usually indicated by a different layout, e.g. bold, italic, or spanning the columns of a newspaper article in the same way as the main title. Nevertheless, automatic detection may be tuned to individual newspapers to reach satisfying results.

5.4.8 Copyright note

Synonyms
By-line, copyright statement

Definition
The copyright statement indicates who is the source of information and therefore responsible for the content of a news article. From a historical point of view, it is interesting to see that the author information becomes more and more important: Whereas for several hundred years news articles did not carry any individual copyright statement, short acronyms for free lancers and photographers were introduced in the 20th century. Nowadays the full name of the author is usually mentioned for every news article of a newspaper. Already in the 19th century copyright notes appear for entertainment, such as novels, poems, cartoons.

Functional value
The value is to inform the reader about the creator of a piece of information or entertainment. This goes together with the increased importance of individual authors as the leading voices of a newspaper.
The copyright statement may be used to increase the quality of the metadata but also for information and retrieval purposes. It might be interesting to see for humanities scholars (even if only the source of information or an abbreviation of a name is available) which news articles are stemming from whom or are written by a specific person.
Again, for automated processing copyright notes can be regarded as embedded particles and are somehow disturbing the running text which may be subject to automated processing.

Automated capturing
We can observe rather strict rules for each newspaper how to handle such copyright notes. In the case of individual authors one can expect that a 'von' or 'by' is used as suffix of this note. The chance to detect this structural element with NLP means is therefore good.

Schema
The party holding the legal copyright to the CreativeWork.
https://schema.org/copyrightHolder

Examples
From 'Die Presse', 13. Mai 1905. Professor Dr. R. v. Wettstein is mentioned as the author of an article about the German School Association. A personal opinion is expressed, as in every Saturday edition at that time (Figure 16).

Nr. 14626.

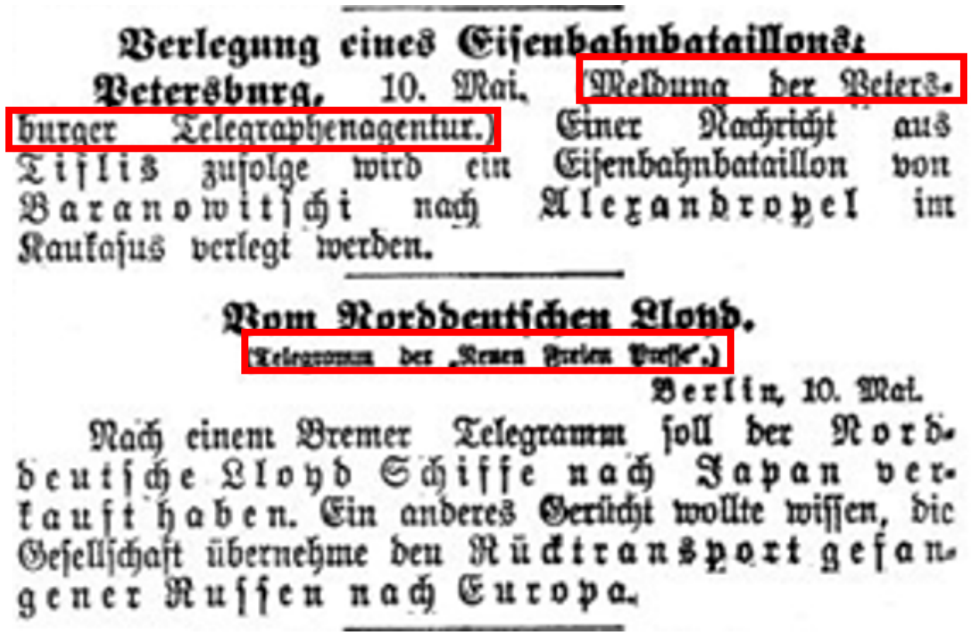
Das Jubiläum des Deutschen Schulvereins.

Von Professor Dr. R. v. Wettstein.

Der Deutsche Schulverein, der größte und älteste der deutschen Schulvereine Oesterreichs, begeht am heutigen Tage die Feier seines fünfundzwanzigjährigen Bestandes. Fünfundzwanzig Jahre ununterbrochenen nationalen Kampfes liegen hinter uns; unwillkürlich drängt die Feier

Figure 16: Copyright note, 1905

In Figure 17 we find two articles from the same issue, where the copyright statement has a slightly different but nevertheless comparable function and expresses the source of information.



Verlegung eines Eisenbahnbataillons.

Petersburg, 10. Mai. (Meldung der Petersburger Telegraphenagentur.) Einer Nachricht aus Tiflis zufolge wird ein Eisenbahnbataillon von Baranowitschi nach Alexandropel im Kaukasus verlegt werden.

Vom Norddeutschen Lloyd.

(Telegramm der „Neuen Freien Presse".)

Berlin, 10. Mai.

Nach einem Bremer Telegramm soll der Norddeutsche Lloyd Schiffe nach Japan verkauft haben. Ein anderes Gerücht wollte wissen, die Gesellschaft übernehme den Rücktransport gefangener Russen nach Europa.

Figure 17: Examples of copyright notes, Die Presse, 1905

'Meldung der Petersburger Telegraphenagentur' indicates that this news was taken from a Russian news agency, whereas the other news was investigated directly by the team of the newspaper editor (Telegramm der 'Neuen Freien Presse'.)

5.4.9 Coverage note spatial

Synonyms
Place name

Definition
News articles commonly indicate the location of the message or story right at the beginning of the item. Coverage notes are defining the location or date of a specific news and should therefore be distinguished from section headings which are indicating just in general the place of origin of the following news items.

Functional value
The user is quickly informed from where a specific news comes and has therefore in many cases some pre-knowledge on the background of the story. The value for document analysis is high since this information can be found nearly from the very beginning of newspaper publishing until today. Surprisingly it has to our knowledge not been used so far for systematic information extraction.
From the historical point of view coverage notes spatial are very similar to section headings which are consisting just of the place name of a specific section.

Automated capturing
Due to the fact that coverage notes follow strict rules within one newspaper and since their repertoire is rather limited they can be automatically detected and regarded as an excellent (indirect) means to find news items automatically.

Schema
The spatialCoverage of a CreativeWork indicates the place(s) which are the focus of the content. It is a subproperty of contentLocation intended primarily for more technical and detailed materials. https://schema.org/spatialCoverage

Examples
Figure 18 shows an example which was taken from the newspaper Gazeta Lwowska (Lemberger Zeitung / Poland) from 16. February 1821. The complete coverage note includes the place and detailed date of the event: Madrytu, and the 11th of January.

5.4.10 Coverage note temporal

Synonyms
Date, dateline

Definition
The exact data which is mentioned at the beginning of a newspaper article and most often part of a general coverage note which also mentions the place of the news. Coverage notes temporal appear already in the 18th century and can be found until the middle of the 20th century in very similar formats.

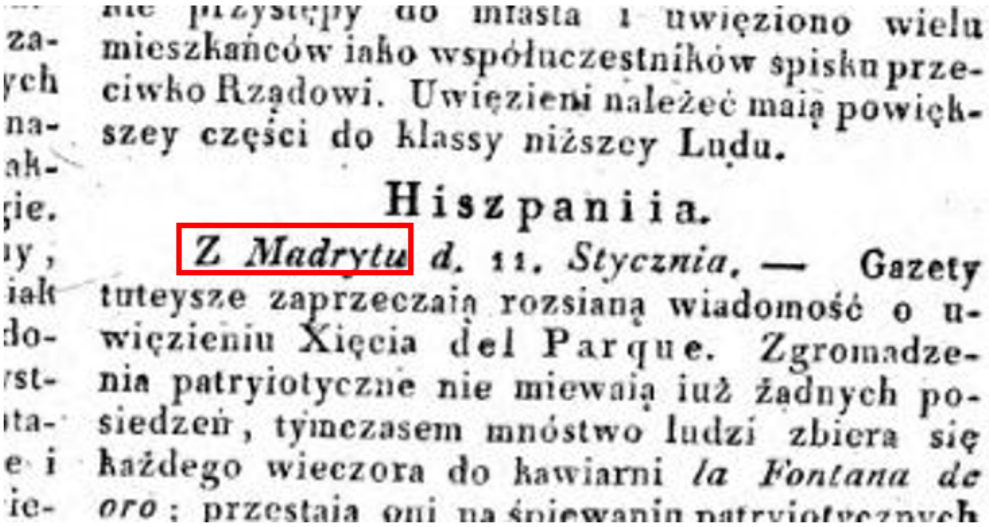

za- mieszkańców iako współuczestników spisku prze-
ych ciwko Rządowi. Uwięzieni należeć maią powięk-
na- szey części do klassy niższey Ludu.
ak-
gie. **Hiszpaniia.**
ıy, Z *Madrytu* d. 11. *Stycznia*, — Gazety
iak tuteysze zaprzeczaią rozsianą wiadomość o u-
do- więzieniu Xięcia del Parque. Zgromadze-
rst- nia patryiotyczne nie miewaią iuż żadnych po-
ıta- siedzeń, tymczasem mnóstwo ludzi zbiera się
e i każdego wieczora do kawiarni *la Fontana de*
ie- *oro*; przestaią oni na śpiewaniu patryiotycznych

Figure 18: Example of a coverage note, 1821

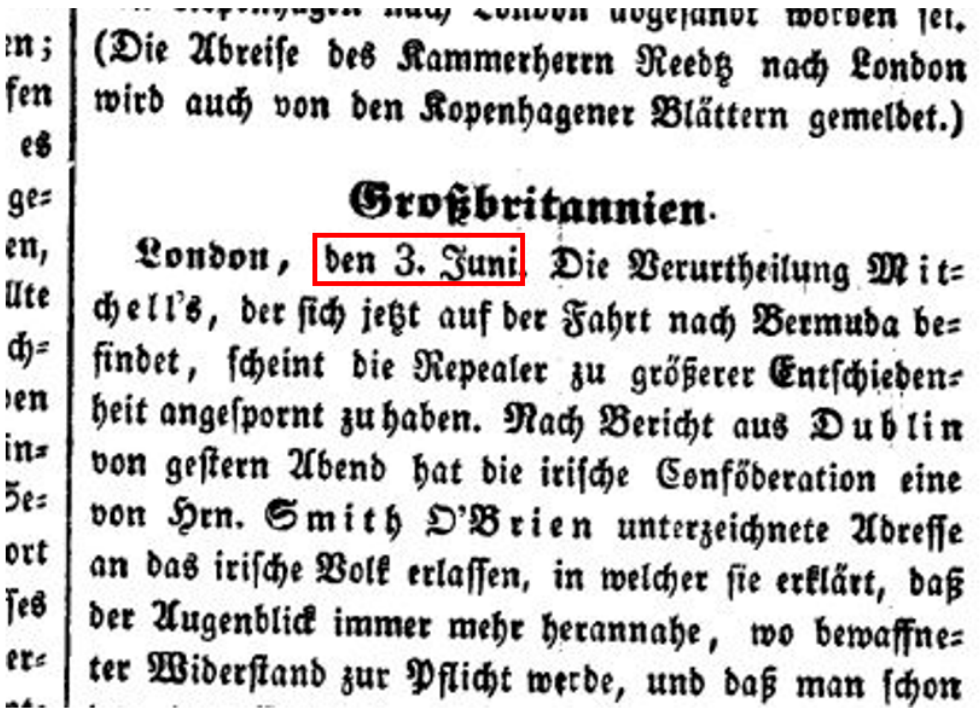
Functional value
With the coverage note temporal the reader gets a detailed information about the 'age' of a news. This was especially important at former times when the transportation of news took several days or – if they came from abroad – even longer. In contrast to the coverage note spatial the information value may not be high in daily newspapers, but in newspapers which are edited only once or twice a week or in irregular intervals an exact date can be more important. Ironically the coverage note temporal became extremely important within social networks - e.g. a Twitter news about an ongoing event may be outdated within minutes.

Automated capturing
Due to the fact that coverage notes follow strict rules within one newspaper and since their repertoire (numbers, days, months) is very limited, they can be automatically detected. In an indirect way coverage notes (both spatial and temporal) could be used to detect the start of a news item since their repertoire is very limited and can therefore be matched against a list of options (place names, dates). The same is true for copyright notes, which also show a very limited number of textual variants.

Schema
The temporalCoverage of a CreativeWork indicates the period that the content applies to, i.e. that it describes, either as a DateTime or as a textual string indicating a time period in ISO 8601 time interval format.
https://schema.org/temporalCoverage

Examples
The example of Figure 19 is taken from the Wiener Zeitung, 9. June 1848. The complete coverage note is: London, den 3. Juni (= London, 3rd June). In this example we can also see that section headings (in this case Großbritannien, Great Britain) and coverage notes

have a very similar origin and are highly related to each other.



(Die Abreiſe des Kammerherrn Reedtz nach London wird auch von den Kopenhagener Blättern gemeldet.)

Großbritannien.

London, den 3. Juni. Die Verurtheilung Mitchell's, der ſich jetzt auf der Fahrt nach Bermuda befindet, ſcheint die Repealer zu größerer Entſchiedenheit angeſpornt zu haben. Nach Bericht aus Dublin von geſtern Abend hat die iriſche Conföderation eine von Hrn. Smith O'Brien unterzeichnete Adreſſe an das iriſche Volk erlaſſen, in welcher ſie erklärt, daß der Augenblick immer mehr herannahe, wo bewaffneter Widerſtand zur Pflicht werde, und daß man ſchon

Figure 19: An example of a coverage note, 1848

5.4.11 Paragraph

Definition
A paragraph is the default unit of a running text and usually provides a single thought or distinct piece or section of a news item.

Functional value
The functional value of paragraphs is to structure a longer piece of text into suitable sections. Usually a paragraph covers one thought or one distinct part of the narrative or message. The value is high, since paragraphs can be seen as a basic building block of any text. The correct separation of the text into paragraphs is important for any kind of natural language processing.

Automated capturing
The automated caption of paragraphs leads to good results within larger units of running texts. Nevertheless, the distinction between paragraphs and short articles might be problematic in many cases. In the case of advertisements and classified advertisements the usage of paragraphs is often arbitrary, respectively real sentences will not be found in the news item.

5.4.12 Illustration (photograph/picture/chart)

Definition
In an illustration the main content is expressed in a non-textual, graphical way. Typical graphical elements are photos, pictures, cartoons, charts, etc.

Functional value
Illustrations are supporting the textual message of a news item. With the development of printing technologies, we see a significant increase of illustrations within newspapers especially in the 20th century.

Automated capturing
As already mentioned above illustrations can be captured with good results in an automated or semi-automated way even on a very basic level.

Schema
An image of the item. This can be a URL or a fully described ImageObject.
https://schema.org/image

5.4.13 Table

Definition
A set of facts or figures systematically displayed, especially in columns and rows. Tables can be found frequently in newspapers, e.g. for stock exchange rates, or railway roadmaps.

Functional value
Tables offer to the reader a comprehensive overview of information which would otherwise be hard to explain with pure narrative means. The value is therefore high.

Automated capturing
Tables can in general be detected automatically but the detailed allocation of facts to rows and columns and their logical order is a serious challenge and a research field on its own. Nevertheless, since some tables appear in a very similar way over years and decades in a newspaper (such as stock exchange rates) it might be possible to take benefit of this fact.

Schema
A table on a Web page. **Note:** Even though the property is defined for tables on web pages the annotation can be utilized to help search engines identify the tables
https://schema.org/Table

5.4.14 List

Definition
A list is a number of connected items printed consecutively, typically one below the other.

Functional value
The value may be high if the items of a list are taken as starting point to match it with external information. E.g. the ranked list of review books may be linked to corresponding resources, such as a library catalogue.

Automated capturing
Similar to tables the automated detection and extraction of fine-grained information from lists is a sophisticated task and can probably only done for very specific newspapers and sections.

Schema
A list of items of any sort.
https://schema.org/ItemList

5.4.15 Continuation note

Definition
One or more words which explicitly indicate that an article is continued on another page or in another issue. Often continuation notes appear on the title page of an issue.

Functional value
The continuation note itself is – as similar structural elements – disturbing the running text of a news item and therefore 'noise'. Nonetheless the functional value of the note is rather high, since it will link together dislocated pieces of a news item.

Automated capturing
Automatic detection of continuation notes is difficult since newspapers handle them very individually. Nevertheless, within any given newspaper the same text phrases are always used to indicate a continuation.

5.4.16 Summary

Synonyms
Preview, billboard

Definition
A summary of the content of a newspaper issue. Often provided as an extra section or on the last page. We include also billboards and previews under this category since they have a very similar purpose.

Functional value
With the increasing complexity of newspapers, it became convenient to provide already a

preview or summary of the content at special sections. Again, the main value is to guide the user in navigating through a complex newspaper issue. The value in terms of information structuring is rather low: content is either referenced or repeated and therefore we can regard these pieces mainly as 'emphasizer' that this content was regarded to be important, but no additional facts or figures are provided with regard to the main content which is presented at another place in the newspaper issue.

Automated capturing
Summaries, previews and billboards usually appear as sections with repeated headings and at well-defined places. They can therefore be handled in the same way as sections. For previews to single articles usually a 'continuation note' is included as well which may be used to identify the summary.

5.4.17 Verbatim quote

Definition
An explicit record of someone's (verbal) expressions. In most cases explicit quotations come with a quotation mark. Verbatim quotes were often part of historical newspapers, mainly when speeches or proclamations of high political representatives are cited within the newspaper.

Functional value
In order to emphasize the importance of the message of a person or an institution a piece of text was included as a verbatim quote. They are usually marked with their layout.

Automated capturing
Verbatim quotes are very specifically handled by each newspaper editor and are therefore hard to detect and to extract in a generic way. Nevertheless, for some newspapers this might be possible on the basis of very distinct rules.

Schema
A quotation. Often but not necessarily from some written work, attributable to a real world author and - if associated with a fictional character - to any fictional Person. Use isBasedOn to link to source/origin.
https://pending.schema.org/Quotation

Examples
A good example of a verbatim quote as a regular means to structure content within a newspaper can be found e.g. in 'Die Presse' from February, 1890 shown in Figure 20. Here the address of the German emperor Wilhelm II at the German Reichstag is cited and indicated in the layout with quotation marks and a smaller font size compared to the default font size of the running text.

Vertreter des fanatischen Araberthums sein.

Die Ansprache Kaiser Wilhelm's an den Staatsrath.

(Telegramm des Correspondenz-Bureau.)

Berlin, 14. Februar. Bei der Eröffnung des Staatsraths trug der Kaiser die Garde-du-Corps-Uniform. Zu seiner Rechten befanden sich Fürst Bismarck und die Staatsminister und zu seiner Linken der Staatssecretär des Staatsraths, Bosse, dessen Stellvertreter geheimer Legationsrath Kaiser und die Generalität. Die Mitglieder des Staatsraths nahmen beliebig Platz. Die Anrede des Kaisers lautete:

„Meine Herren Mitglieder des Staatsraths! Durch meinen Erlaß vom 4. d. sind Sie davon unterrichtet worden, daß es mein Wille ist, das Gutachten des Staatsraths über diejenigen Maßnahmen zu hören, welche zur besseren Regelung der Verhältnisse des Arbeiterstandes erforderlich sind. Es entspricht der Bedeutung, welche der Staatsrath in der Monarchie einnimmt, daß die wichtigen, auf diesem Gebiete einer gedeihlichen Lösung harrenden Fragen von Ihnen einer gründlichen Erwägung unterzogen werden, bevor die aufzustellenden Gesetzentwürfe an die parlamentarischen Körperschaften gelangen, denen die endgiltige Beschlußfassung darüber verfassungsmäßig zusteht.

Figure 20: Verbatim note, indicated by smaller font size, 1890

6 News items and structural elements within NewsEye

In order to simplify the task we decided to tackle not all structural elements (as defined in Section 5.4) but to make a meaningful selection. So we agreed on a smaller subset of these structural elements used for the first machine learning approach getting articles/news items inclusively some structural elements. However, getting meaningful news items automatically would be already a big success story at the current state of research.

6.1 Structural element list of NewsEye

- headings (at the beginning no differentiation between headings and sub-headings)
- paragraphs
- illustrations
- captions
- lists
- tables

In the training data all these structural elements are marked and in addition the news items were annotated. For news items we distinguish three different types: articles, advertisements and classified advertisements. The difference of advertisements and classified advertisements

can be defined as follows: advertisements mainly want to sell something very specific in a rather strong manner, products, etc. Classified advertisements want to inform and in a second step to sell, e.g. official announcements, open jobs, houses to rent, etc. For researchers it makes of course a lot of sense to know of what type a news item is. And with the help of a filter it is easily possible to restrict searches to one type only.
As additional elements we decided to also mark horizontal and vertical separators. This can be done in an automatic way by using the OCR engine Abbyy Finereader. Separators can be used as an additional feature for the AI methods helping them recognising the document structure.
With these defined elements the generation of training data was started in Y2.

7 Data schemes and their integration into the NewsEye Demonstrator

Because of the number of documents available and the quantity of data, it is not reliable to count on querying the pure full text in these documents. For users to be able to find information effectively and as quickly as possible, the content of these documents must be indexed.
The index used by the NewsEye Demonstrator is Solr. This tool is at the base of a lot of search engine, as it allows for quick retrieval and querying of documents. Those documents can be retrieved in various formats (XML, JSON, CSV, etc).

7.1 PAGE XML documents in SOLR

In the NewsEye project we use PageXML documents and METS/ALTO documents which represent entire issues, their pages and the separated articles/news items. The PAGE XMLs from Transkribus - enriched with improved text recognition and article separation - get converted into SOLR documents and in this section we show examples of the internal representation of those documents.

7.1.1 SOLR example of an newspaper issue


```
{
    "id":"aze19000816",
    "has_model_ssim":["Issue"],
    "title_ssi":"Arbeiter Zeitung 1900-08-16",
    "date_created_ssi":"1900-08-16-",
    "date_created_dtsi":"1900-08-16T00:00:00Z",
    "language_ssi":"de",
    "original_uri_ss":"http://anno.onb.ac.at/cgi-content/anno?aid=aze&datum=19000816",
    "nb_pages_isi":6,
    "thumbnail_url_ss":"http://localhost:3000/iiif/aze19000816_page_1/full/,200/0/default.jpg",
    "member_ids_ssim":["aze19000816_page_1",
        "aze19000816_page_2",
        "aze19000816_page_3",
```

```
        "aze19000816_page_4",
        "aze19000816_page_5",
        "aze19000816_page_6"],
    "year_isi":1900,
    "member_of_collection_ids_ssim":["arbeiter_zeitung"],
    "all_text_tde_siv":"Preis für die Provinz\n10 hel..."
}
```


7.1.2 SOLR example of a newspaper page


```
{
    "id":"aze19000816_page_1",
    "has_model_ssim":["PageFileSet"],
    "page_number_isi":1,
    "width_isi":3465,
    "height_isi":5359,
    "mime_type_ssi":"image/jpeg",
    "iiif_url_ss":"http://localhost:3000/iiif/aze19000816_page_1"
}
```


7.1.3 SOLR example of an newspaper article


```
{
        "id":"aze19000816_article_68",
        "language_ssi":"de",
        "all_text_tde_siv":"Abgeordneten Kiesewetter vertreten...",
        "date_created_ssi":"1900-08-16",
        "date_created_dtsi":"1900-08-16T00:00:00Z",
        "level":"0.articles",
        "year_isi":1900,
        "from_issue_ssi":"aze19000816",
        "member_of_collection_ids_ssim":["arbeiter_zeitung"],
        "canvases_parts_ssm":["http://localhost:3000/iiif/aze19000816/canvas/
                page_3#xywh=2289,360,1045,48",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2289,408,1045,49",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2289,450,1045,41",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2289,487,1045,60",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2289,539,1041,60",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2289,577,1041,64",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2289,618,1041,56",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2289,663,1041,56",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2285,708,1049,49",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2289,753,1041,41",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2289,790,1041,53",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2289,835,1041,57",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2285,880,1038,53",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2289,925,1041,53",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2285,963,1045,52",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2289,1008,1041,60",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2289,1053,1041,75",
          "http://localhost:3000/iiif/aze19000816/canvas/page_3#xywh=2285,1094,195,67"],
        "thumbnail_url_ss":"http://localhost:3000/iiif/
                aze19000816_page_3/2285,360,1049,801/!400,200/0/default.jpg",
```

```
        "has_model_ssim":["Article"]
}
```


7.2 Named entities in SOLR

As for the named entities, which are provided using the IOB format, a mapping needs to be established between the original text and the tokens available in the IOB file. This allows to precisely locate entities in the text of issues and articles. Moreover, when an entity mention is linked to an entry in a knowledge base, this entry needs to be indexed as well to display the associated label in the demonstrator interface. Here is an example of the internal representation of a linked entity and of an entity mention.

7.2.1 SOLR example of a linked entity


```
{
        "id":"entity_LOC_Q90",
        "entity_type_ssi":"LOC",
        "kb_url_ssi":"https://www.wikidata.org/wiki/Q90",
        "label_en_ssi":"Paris",
        "label_fr_ssi":"Paris",
        "label_fi_ssi":"Pariisi",
        "label_sv_ssi":"Paris",
        "label_de_ssi":"Paris"
}
```


7.2.2 SOLR example of an entity mention


```
{
    "id":"entity_mention_uusi_suometar_9739_1",
    "linked_entity_ssi":"entity_LOC_Q1278356",
    "issue_id_ssi":"uusi_suometar_9739",
    "type_ssi":"LOC",
    "article_id_ssi":"uusi_suometar_9739_article_9",
    "mention_ssi":"Porin",
    "issue_index_start_isi":1512,
    "issue_index_end_isi":1517,
    "article_index_start_isi":507,
    "article_index_end_isi":512,
    "stance_fsi":0.0
}
```


8 Conclusion

In this deliverable we identified the different aspects of the 'data models' in the lifetime of the NewsEye project and beyond; What there is, what the ideal data mode would be and how can we reach the goals of the project using them. Since each of the partner institutions and research groups already use their own data models we tried to combine and respect all these existent working models as well as to integrate new ones. Moreover we converted one format into another several times to allow the usage of existing interfaces, workflows and

tools. We showed different ways to represent and exchange data.

As a summary, we hereby list the main data formats and models and how they were used in the NewsEye project:

- IIIF: distribution format mainly for images
- METS/PAGE: main export format of Transkribus, contains text, articles and other structural elements
- METS/ALTO: delivery format for the libraries as their systems can already deal with it
- ConLL/IOB: storing NER, NEL and stance training data; can be exported from Transkribus as well and is the input for the training algorithms
- ConLL/IOB: storing NER, NEL and stance production data; is the output from the used NE recognition tools
- JSON: used as input format for SOLR index and contains text as well as articles and NER, NEL and stances. Other formats (e.g. PAGE XML, ConLL/IOB) get converted into JSON which is the import format of the NewsEye Demonstrator.

Additionally to the above data models, this deliverable introduced main concepts within newspapers. All of them will get important in future approaches to detect structural elements in newspapers. At the moment only a small subset was selected to start with.

Another aspect was the presentation of the IIIF data model and the possibilities this offers. We think that this framework provides good solutions for the future but needs to be adapted and extended in one or the other direction. Not all results can be easily mapped into IIIF since some standardization still needs to be discussed and integrated to perhaps become the 'all covering' data model of the future in the digital heritage world.

Finally we tried to give a good picture of how the data of the project arrives, enters and updates the Demonstrator so that it can act as the final distributor of all the NewsEye results.

References

[1] Axel Jean-Caurant and Antoine Doucet. "Accessing and Investigating Large Collections of Historical Newspapers with the NewsEye Platform". In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. JCDL '20. Virtual Event, China: Association for Computing Machinery, 2020, pp. 531–532. ISBN: 9781450375856. URL: https://doi.org/10.1145/3383583.3398627.

[2] Stefan Pletschacher and Apostolos Antonacopoulos. "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework". In: *2010 20th International Conference on Pattern Recognition*. IEEE, Aug. 2010. DOI: 10.1109/icpr.2010.72.

[3] Jacob Eisenstein. *Introduction to Natural Language Processing (Adaptive Computation and Machine Learning series)*. The MIT Press, Oct. 2019. ISBN: 0262042843. URL: https://www.xarg.org/ref/a/0262042843/.

[4] Erik F. Tjong Kim Sang and Fien De Meulder. "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 2003, pp. 142–147. URL: https://www.aclweb.org/anthology/W03-0419.

[5] Tom Crane. *Looking up and looking down: IIIF Resources, Intellectual Objects, and Units of Distribution*. Digirati. Dec. 2017. URL: https://resources.digirati.com/iiif/an-introduction-to-iiif/looking-up-and-down.html.

[6] Gene Loh. "Linked data and IIIF: Integrating taxonomy management with image annotation". In: *2017 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)* (2017), pp. 50–55.

[7] Tobias Strauss. "Decoding the Ouput of Neural Networks – A Discriminative Approach". PhD thesis. University of Rostock, 2016.