DELIVERABLE

Project Acronym: Europeana Newspapers

Grant Agreement number: 297380

Project Title: A Gateway to European Newspapers Online

D5.2 Europeana Newspapers
METS ALTO Profile (ENMAP) – External Release - DRAFT

Revision: 1.0

Authors: **Günter Mühlberger (UIBK)**

Contributions: **All partners from WP5 Metadata**

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	x
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
D5.1 – 2	15/01/2013	Günter Mühlberger et al.	UIBK	Update after internal discussions
D5.1– 3	e01/2013	Günter Mühlberger et al.	UIBK	Minor changes
D5.1 – 4	02/2013	Günter Mühlberger et al.	UIBK	Minor changes
D5.2 - 1	06/2013	Günter Mühlberger et al.	UIBK	New version
D5.2 – 2	09/2013	Günter Mühlberger et al.	UIBK	Minor changes
D5.2 – 3	12/2013	Günter Mühlberger et al.	UIBK	Minor changes

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Contents

1. Executive Summary

Work package 5 “Metadata Best Practice Recommendations” runs from M1 to M36 and has two main objectives:

- The **first objective** is to provide a metadata schema that fits the requirements of the Europeana Newspaper Project (ENP). These requirements are derived one the one hand from the enhancement tasks carried out in work package 2 “Refinement of digitized newspapers” where 8 million newspaper pages are OCR process and another 2 million pages are segmented into smaller units such as articles or sections. On the other hand it was necessary to set up a unified format in order to deliver all these files to The European Library (TEL) for further integration into their Newspaper Content Viewer which is developed as part of work package 4 “Aggregation and presentation of digitized newspapers for Europeana”. This first objective was realized as part of task 5.1. in M12 with deliverable **D5.1. First internal release of the format with initial online resource for documentation.**

- The **second objective** is to take this internal format, to further enhance it and to make it available to the public as best practise recommendations. This process consists of two steps: First of all it is planned to provide an initial release (which was planned for M18 but was now postponed to M24 and which is represented with this paper) and secondly to gather feedback from libraries and stakeholders and to adapt the format according to their feedback until the end of the project. The final release is therefore planned for M36. In addition to the foreseen online documentation (example files, guidelines and decision trees) and exceeding the original work plan also several tools have been developed in order to generate and process the best practise format.

The suggested information package conforms with the OAIS (Open Archival Information System) and is implemented as a METS (Metadata Encoding and Transmission Standard) container. Data stemming from OCR (Optical Character Recognition) processes are stored within ALTO (Analyzed Layout and Textual Object) files. With the release of deliverable D5.2 a special focus has been laid on the so-called “Structural Map”, i.e. the internal structure of newspapers. Since this field has long been underexposed within the libraries world we believe that our recommendations will find some attention within the community. The working name for our information package is **ENMAP** which stands for: **E**uropean **N**ewspaper **M**ETS **A**LTO **P**rofile.

A first internal release of ENP-ENMAP took place right before the Innsbruck Meeting (September 2012), followed by an internal feedback cycle. Updates were included during and after the Meeting in The Hague (November 2012). From January 2013 onwards ENMAP (simple) served as the internal project format and several millions of files where processed according to this project standard.

Already at the project meeting Paris (February 2013) first discussions started concerning the public release of ENMAP. The most important addition was to extend it with a detailed description of the structural elements appearing in (historical) newspapers. This issue was also brought up in the Belgrade workshop (June 2013), and at the Amsterdam workshop (September 2013). It turned out that this task required more discussion and more adaptation and clarification work than we had expected. Several working papers were exchanged and also dedicated meetings took place for finding a common ground. Highly valuable input came especially from CCS GmbH and partners from work package 5 of the project.

2. ENMAP (simple)

2.1 Design considerations

As indicated in the executive summary we distinguish between this first internal release of the Europeana Newspaper METS ALTO profile, called ENMAP (simple) and the external release in M1which aims to become a best practise model for European libraries. ENMAP (simple) is intended to provide a simple but effective encoding for all newspapers that are refined within the Europeana Newspaper Project (ENP). The main purpose is therefore to create a Submission Information Package (SIP) for the TEL Newspaper Viewer which will be developed in work package 4 “Aggregation”.

In contrast to this project based approach this external release will go one step further and also include mainly questions concerned with the structural order within newspapers. Though there would be some overlapping we have excluded the question of born-digital newspapers since it would on the one hand significantly increase the complexity as well as not be covered by the objectives of the Europeana Newspaper Project.

2.2 Implementation principles

2.2.1 Metadata Encoding and Transmission Standard (METS) as container format

Though there would be other formats as well such as MPEG21 it is commonly agreed among libraries that a METS container is the preferred way to encode information from digitised newspapers. METS is hosted by the Library of Congress. We will use the latest version 1.91.

2.2.2 MODS as container for descriptive metadata

Though MODS (Metadata Object Description Schema) does not have the same widespread use as MARC21 the natural connection between MARC21 and MODS is the strongest argument for choosing MODS. Most libraries use MARC (or a dialect of it) and are able to convert MARC into MODS. All descriptive metadata will therefore appear in the MODS format. MODS is also hosted and maintained by the Library of Congress and we will use MODS 3.4.

2.2.3 MIX as container for technical metadata

Technical metadata can be stored either within the content files itself (e.g. tiff tags or EXIF tags) or explicitly in the METS container. Nevertheless, from an implementation point of view, we decided that some technical metadata which are implicitly available in the content files are extracted and stored in the METS administrative section. We will therefore use the MIX format to store the most important technical metadata, such as format, width, height, size, etc.

NISO Metadata for Images in XML (NISO MIX) is also hosted and maintained by the Library of Congress and will be used as version 2.0.

2.2.4 PREMIS (Preservation Metadata Maintenance Activity)

PREMIS offers a sophisticated model and data dictionary to store information which might be relevant in the digital life cycle. It can be seen as a must for long-term digital preservation applications and for the design of an Archival Information Package (AIP). For this reason PREMIS is recommended for preservation purposes, but for the purposes within ENP we will not use the PREMIS format since the main functional requirement is to provide a SIP for the TEL Content Viewer.

2.2.5 ALTO (Analysed Layout and Text Object) as container for textual data

There are many ways to store text that is connected to an image which has been OCR processed: simple text files, PDF files with a text layer, DjVu Files with a text layer, proprietary XML vendor files, etc.

Nevertheless ALTO is the only format which has been created within the libraries community and which is widely used and also supported by industry (ABBYY). ALTO will therefore be the preferred format for OCR data. It is hosted and maintained by the Library of Congress and will be used as version 2.0.

2.2.6 Encoding of Named Entities

Based on the ALTO files some textual material will be enhanced with Named Entities in work package 2 “Refinement of digitized newspapers”. A discussion on this issue started already in spring 2012. Since there are no standard formats available further investigations need to be carried out. From the point of view of the project work plan there is no strong dependency. Most likely a separate XML file will be used that contains Named Entities and links to the coordinates according to the information set out in the ALTO files. This issue will be clarified in the remaining project duration.

2.2.7 Rights metadata

Though rights are an important issue it has to be accepted that – e.g. in contrast to MARC, MODS, METS, ALTO, etc. – the discussion on formats and models is still fluid and any decision is highly arbitrary.

In order to avoid complexity we assume that the rights information is kept outside the ENMAP information package. Only an “owner” tag is kept in the METS header section. The “owner” (which is within the project in all cases the library that provides the data) has the full right on the information package (whatever this may include in detail).

2.2.8 Naming of directories and files (for the file section of ENP-ENMAP)

Naming of directories and files is important since it has an impact on the delivery procedure within the project. In order to avoid uncontrolled exchange of data a strict system was set up that also contains some rules on the naming of directories and files. The detailed description can be found in WP2 Specification papers delivered in the first project year.

The most important prerequisites are that all newspapers (1) must reside in a directory structure that contains on the one hand a unique identifier as well as (2) the date of a newspaper issue. This means also that content providers are free to name their files and issues as long as the two requirements from above are fulfilled.

2.2.9 Issue centric

The issue is the natural physical and logical entity of a newspaper. The digital object of an issue must contain all the information which is necessary to understand the content for human beings as well as make it readable for machines. This means that we will produce METS files on issue level, but not on newspaper level. The newspaper level is covered by the corresponding MODS and MARC21 records.

2.2.10 Editions

We speak of editions of a newspaper if a title is published several times a day, either as temporal edition (typically: morning-evening) or spatial (regional edition).

Though METS would allow to cover this process it would increase complexity significantly. Therefore we opt for a simpler solution:

If a library has collected a separate edition in a separate way, or if during the scanning process the edition is handled as separate run, than an extra METS file is generated where either another identifier is used for the respective MARC record, or where the title information within MODS keeps the data on edition level. But there will be no linking between the two METS files.

If a library does not distinguish between editions, e.g. because the bound volume contains all editions and during the scanning process no separation was carried out, the information that the METS file keeps several editions is also available via the link to the respective MARC record and via the title information of the MODS descriptive metadata part within METS but this does not have any consequence in the organisation of the METS file.

2.2.11 Supplements

The same solution is suggested for supplements: If a supplement is treated separately also separate METS records will be generated, if it is handled as part of the newspaper the descriptive metadata of the "parent" newspaper should keep the information that also a supplement is included.

So there is no linking between the METS files of several editions and supplements.

3. ENMAP (full)

3.1 General considerations

Though newspapers are among the most frequently used materials in libraries, their recording in a library catalogue is rather shallow. Apart from general information about the (changing) titles and (changing) publishers of newspapers as well as missing volumes and issues the information in usual library catalogues is tenuous. In rare cases libraries have started to record "important articles" and in documentation centres one may find catalogues with newspaper articles or even clippings of single articles but in general the indexing of newspapers is poor. With the digitization of newspapers this situation changes dramatically. Now automated processes are possible (Optical Character Recognition, Optical Layout Recognition) that provide on the one hand a full-text index for information retrieval, as well as a segmentation of newspaper pages into physical (blocks, lines, strings) and logical (headlines, articles, by-lines, captions) units.

Therefore the need for a common understanding of "what can be found in a (historical) newspaper" and "how can it be named" is of eminent importance. One can say that only with the digitisation of large amounts of newspapers libraries for the first time face the problem of how to structure a newspaper in a meaningful and effective way? How deep shall the indexing go, who shall do it, can it be done automatically, and what would be the benefits?

This paper will not be able to provide exhaustive answers to these questions but it is a first attempt to summarize the discussion on this topic and to provide a comprehensive proposition how these questions could be answered.

This data dictionary has the ambition to provide a classification system for the content of (historical) newspapers in order to encourage standardisation in this domain. It offers criteria and definitions how to structure a newspaper into its components. In the ideal case one might open a newspaper and be able to classify every single item according to the guidelines set up in this paper.

As with all classifications there are several difficulties connected: Classifications are somehow arbitrary and depend on the applied criteria. There is certainly more than one view on a complex thing like a newspaper and therefore other approaches would be justified as well. In addition there are also different conventions and traditions depending on the country of publication, e.g. newspapers in Great Britain will have a different repertoire of structural elements compared to those in Germany or Poland and current newspapers will look different to newspapers from the 18th century.

Moreover the distinction between all the classes may be hard and the definitions used may not be applicable in all situations. As a matter of fact reality is always more complex and offers many surprises so that the best classification scheme will not be able to cover every single case.

This said we believe that with this data dictionary and with the criteria and examples which are provided in this paper an overwhelming majority of articles and elements in European newspapers can be classified in a transparent and fairly objective way.

We also believe that this can be an important contribution to a standardisation for the description of historical newspapers which would be one of the prerequisites for a pan-European newspaper service.

3.2 Main concepts

A newspaper is a collection of diverse materials: News articles, job offers, advertisements, pictures, stock exchange charts, obituaries, letters to the editor, serial novels, cartoons, weather forecasts, marriage notes and many, many other content pieces can be found. Often these discrete pieces are put together in sections, such as “Foreign affairs” or “Obituaries” or simply “Latest news”.

In order to make the newspaper attractive to readers, newspaper publishers also developed a number of features for the layout of the paper, e.g. headlines, sub-titles, by-lines, or captions. Since a typical newspaper is published on several days of a week the structure is repeated rather often. Sections, such as “Foreign affairs” or “Obituaries” frequently occur every day more or less at the same “location” within an issue.

Since the beginning of newspapers in the 17th and 18th century newspapers changed their appearance strongly and their repertoire concerning their internal structure differs significantly from today. But from a historical perspective it seems that the history of newspapers is mainly a history of “differentiation” which means that new concepts were added but old concepts very rarely were completely given up.

Also today we “understand” the structural “behaviour” of historical newspapers. It is exactly this “intuitive understanding” which shall be made explicit with the considerations provided in this paper.

From our point of view three main concepts are sufficient to describe the structure of newspapers. These three main concepts are:

- Newspaper Content Items (NCI)
- Newspaper Structural Elements (NSE)
- Newspaper Content Sections (NCS)

The following figure illustrates these three main concepts:

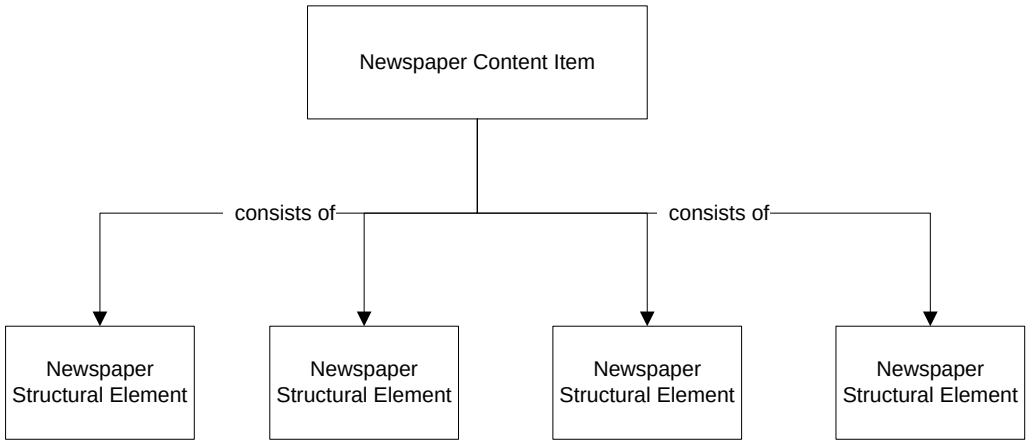


Fig. 1 Newspaper Content Items and Newspaper Structural Elements

If we take a newspaper article or a book review as examples: Both are “building blocks” or “components” of the content of a newspaper. Both may consist of a title, a sub-title, a byline and in case of the book review probably a reference to the reviewed book. These structural elements are the components of the content item. In contrast to the newspaper content items, which are defined mainly by their contextual message, newspaper structural elements are defined by their functionality within the text.

The situation becomes a bit more complicated due to the fact that Newspaper Content Items are often arranged in larger units, which we will call Newspaper Content Sections. Again two examples: A collection of short articles may appear as “Latest news”, and a section called “Job offers” may consist of a large number of single job announcements.

Note that it is not allowed that Newspaper Content Items contain other Newspaper Content Items. The main reason for this constraint is that this strict rule will make it easier to define Newspaper Content Items and to apply this definition to the actual material.

The following figure provides an overview:

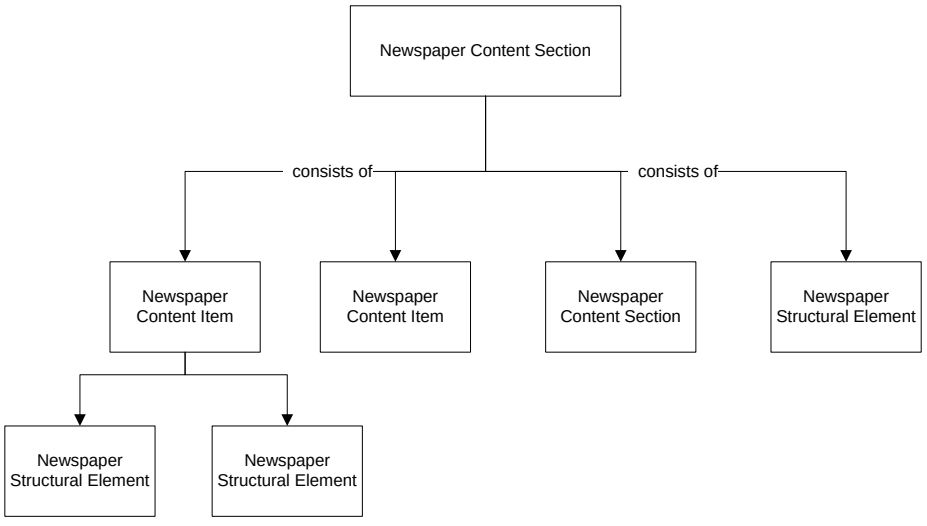


Fig. 2 Newspaper Content Sections

As we can see a Newspaper Content Section may consist of several Structural Elements, e.g. a title, sub-title and an author (responsible person for the section) and it may consist of several (actually at least two) Newspaper Content Items. Note that Newspaper Content Sections must contain at least one Newspaper Content Item.

In the following chapters we will provide a more detailed view on these three main concepts.

4. Newspaper Content Items (NCI)

4.1 General considerations

A “newspaper content item” is a discrete unit of content within a newspaper. NCIs can be understood as a distinct “message” which is sent from the creator to the reader mediated by the physical appearance of this content item within this newspaper.

In the following we provide three criteria which can be used to define the concept in more detail.

4.2 Definition

4.2.1 Distinct message

An NCI typically provides one distinct message. E.g. it is a report about the progress of political negotiations, or about a car accident, or about a crime case at court. Single messages that can be clearly separated from others may also be found in the job announcement section, or in the letters to the editor section. Each single job announcement and each single letter is one NCI. So the focus of an NCI is always the content, even if the content may be a long list of stock exchange rates, or job announcements.

4.2.2 Distinct authors/responsibilities

NCIs are also an intellectual entity in the sense that the “copyright” or the “editorial responsibility” can be clearly named and (often) separated from other pieces. The authorship of a specific NCI may help to distinguish further. In many cases contributors such as authors/journalists, photographers, illustrators or cartoonists, are explicitly marked in the article.

4.2.3 Layout features

In many cases the layout of a newspaper indicates the “borders” between NCIs. E.g. separators are used or the headline indicates the start of a “new” NCI. Note, that the layout is only one criterion among others to classify an NCI. As we have emphasized above, content related criteria are more important.

In short: If someone would re-edit all articles of a famous journalist as a book, he would be interested to keep the content as authentic as possible, but re-format it according to the new media, or in other words: The Newspaper Content Item would be part of the book collection, not the actual representation or manifestation of a given article.

One may argue that such as “content based definition” may be unrealistic to realize in a mass-digitisation project where only limited resources are available. This may be true from a practical viewpoint, but from an overall perspective it is necessary to have a strict criterion available (“discrete unit with a clear message”) in order to guide the work and to be able to make a “final decision” that

is not based on the layout (which may be arbitrary and changing a lot) but on the text and the message itself.

5. Newspaper Content Sections (NCS)

5.1 General considerations

In the chapter above we have argued that a newspaper issue can be seen as a compilation of distinct content items. But as a matter of fact, some of them are tied together and are forming larger units. Such units play an important role in newspapers, since they are structuring the content and they are repeated over longer periods.

An example would be the section “Latest news”, which may appear on one of the last pages and may contain several Newspaper Content Items arranged according to the criterion, that the particular NCI was brought to the newspaper’s attention at a late date. Other typical sections could be “Political affairs”, “Local news”, or “Sports”.

For digitised newspapers such “sections” are highly important since they can (often) easily be detected and named since their headlines are repeated over years and decades and they appear with a similar layout and on similar locations or days within a week. Actually it might not be worth to segment single articles within a section but it will in many cases be sufficient to mark the section as a whole to guide the reader in a convenient way.

Thought we could consider regarding sections also as content items, there are several criteria which allow us to distinguish them from NCIs.

5.2 Definition

5.2.1 Frequently repeated

Newspaper Content Sections are repeated over a period of time, and – in contrast to article series – they are, in principle, never-ending. Often their frequency is based on a strict rhythm, e.g. some sections will appear only in the Friday edition. The fact that they are repeated is the most important distinction to Newspaper Content Items, which are per se unique. Though every newspaper developed its own “vocabulary” of sections these items are rather stable over the years and decades. A list of this vocabulary will in many cases be a good starting point for automated structuring.

5.2.2 Collection of Newspaper Content Items

A Newspaper Content Section is usually a collection of several NCIs. The criteria for the compilation may depend on the actual content (“Foreign affairs”, “Local news”) or on formal parameters (“Letters to the editor”, “Latest news”). In contrast to NCIs they have neither a distinct message, nor are they an intellectual unit. The fact that some NCIs are subsumed under the heading “Job offers” does not implicate that it is the same job being offered, or that all “Latest news” are dealing with one single event. Newspaper Content Sections may be better compared to the functionality of a “subject heading”, or an “indexing term” that specifies an aspect many content items share.

5.2.3 Discrete units

Newspaper Content Sections usually appear with a specific “section heading”. Similar to Newspaper Content Items also Newspaper Content Sections are separated from each other, respectively from other Content Items by the layout. A distinctive headline, or frames and separators usually indicate the start of a particular Newspaper Content Section.

5.2.4 Differentiation from columns

Apart from sections also so-called “columns” can frequently be found in newspapers. These also regularly appear with a section heading that is repeated. Nevertheless we see two distinct criteria to separate them from Newspaper Content Sections: They usually consist of just one Content Item (and have therefore one distinct message) and they have a named author (often the author is more important than the actual article).

The value of Newspaper Content Sections is rather high, since a correct extraction allows classifying articles according to the (historical) categories of a given newspaper. A hit in the full-text of “Local news” will be very much different from a hit in “Job offers”. Text mining and automated classification will benefit considerably from these classes.

6. Newspaper Structural Elements (NSE)

6.1 General considerations

As we have seen Newspaper Content Items and Newspaper Content Sections are mainly defined by criteria related to the contents. Their actual appearance within the newspaper plays only a secondary role, whereas their message and content are primarily used to differentiate them. In order to cover this aspect of newspapers and to exploit it for a deeper understanding we need to take a closer look at the appearance of Newspaper Content Items and Sections.

6.2 Definition

6.2.1 Functional value

Structural elements are defined by their functionality within a content item and within a newspaper. E.g. headlines raise the attention of a reader and inform him or her about the main content of a news article. The copyright note or by-line provides the information, who, where and when an article was written, the caption explains the content of a picture, table or chart, etc. The main function of Newspaper Structural Elements is therefore related to their functionality towards the reader, and aims mainly at supporting him or her in understanding the content and being able to navigate through this complex content easily.

6.2.2 Part of a Content Item or Section

Structural elements do not appear on their own, which means that they are always part of a larger unit, in our case part of a Newspaper Content Item or Newspaper Content Section. The third main concept of our classification scheme is “structural elements”. In contrast to Newspaper Content Items structural elements are not discrete units but always part of a content item. A content item may consist of several structural elements, such as headlines, sub-headlines, leads, pictures, copyright notes, etc. In other cases it may just be a paragraph with some textual mark up (bold, italic, etc.) indicating a headline.

6.2.3 Semantics of structural elements

Due to the fact that the repertoire of structural elements was developed over a long period of time a specific semantic connected to the layout is associated with these. E.g. even if we look at a newspaper from far away, or in a completely foreign language, we will “understand” some of the structural elements, such as headlines, sub-titles, caption lines, etc. It is exactly this aspect that makes structural elements so interesting for automated processing and enhancement via Optical Layout Recognition. But again, we have to emphasize that also in this case the layout plays a secondary role whereas the function of a given element should primarily be taken as distinctive criterion.

7. Other structural features

7.1 Reading order

Though European languages share a default reading order from left-to-right and from top-to-down the actual reading order within a newspaper can be rather complicated and not clear without actually reading the page. Moreover for some elements it makes no sense to define a reading order, such as the running title, page number, a graphical advertisement, etc. since these have different functionalities compared to the running text.

We therefore need a clear distinction between those elements which fall within a reading order and those which are outside the reading order.

The correct reading order is rather important, not so much for full-text retrieval, but for all text mining operations that are dependent on the correct order of sentences. Also for displaying e.g. articles on an e-book reader or for clipping services the reading order is of eminent importance.

7.2 Separators

Separators are graphical elements which are used to structure the layout of a newspaper. Typically we find horizontal and vertical lines between e.g. the running title and the body of text, as well as between content items. But also characters such as *** are used for the same purpose.

The value is only given for automated processing and for understanding the layout of a newspaper. Separators may play a role for keying instructions to service providers to give them some elements at hand for fast manual distinction of content items.

No manual caption of separators is foreseen, but OCR engines are capturing separators automatically and this information can be found in the OCR XML or ALTO file and may be used for automated processing.

8. METS/ALTO Representation

8.1 General

In the following we will describe ENMAP (simple) and ENMAP (full). ENMAP (simple) is designed for digitized newspapers that are structured on issue level but where no further structural tagging is carried out. This means that the single page is the only element within the structural element.

In contrast ENMAP (full) consists of a large number of Content Items, Structural Elements and Content Sections. Their internal organisation is complex and requires a number of decisions. It is obvious that there are many options how to transform this structure into a METS file and that our paper can only provide one solution among many others.

8.2 ENMAP (simple)

8.2.1 Part of a Content Item or Section

All newspapers that are refined in the Europeana Newspaper Project are described in the so-called "Master list" which resides on the project Sharepoint server. It contains two main identifiers:

1. The ENP identifier which is only used for internal project purposes.
2. A so-called "library identifier", which is a MARC21 or similar identifier used to link to the full record of the newspaper e.g. in a library catalogue.

We assume that the definite information on newspaper level is recorded in an external resource, such as a MARC21 record or a respective library catalogue. Note that we do not provide METS files on newspaper level which would complicate the workflow a lot. As described above, the issue is the natural unit of a newspaper and therefore a METS file will be created to describe each single issue of a newspaper.

This section refers to the example issue which we created and have set up on the Sharepoint server. It is a natural part of the online documentation of this report.

8.2.2 METS Root Element and header

Object ID: To uniquely identify the METS object within ENP we use a combination of the internal identifier and the date of the issue.

Type: To distinguish the METS objects we use the "Type" attribute of the root element and provide as type for all files "Newspaper".

An example for a header is provided in below:

```
<mets:metsHdr CREATEDATE="2013-01-09T11:53:35" LASTMODDATE="2013-01-09T11:53:35"
   RECORDSTATUS="SUBMITTED">
   <mets:agent ROLE="OTHER" OTHERROLE="OWNER">
     <mets:name>LFT</mets:name>
   </mets:agent>
   <mets:agent ROLE="CREATOR" TYPE="ORGANIZATION">
```

```
    <mets:name>UIBK</mets:name>
   </mets:agent>
   <mets:agent ROLE="CREATOR" TYPE="OTHER" OTHERTYPE="SOFTWARE">
    <mets:name>ABBYYFineReaderEngine10</mets:name>
   </mets:agent>
  </mets:metsHdr>
```

For the creation and last modified date we use the coordinated universal time (UTC).

Agents: Three agents are used within ENMAP:

(1) Creator of the type “organization”, for recording the agency which created the METS object. In the case of the project this is either CCS or UIBK.
(2) Creator of the type “software” and
(3) Owner of the METS file, which is in our case always the library which provides the data.

8.2.3 METS Descriptive Metadata section – MODS

The descriptive data are coming from the “Master List” and are a condensed version of what we regard to be necessary for supporting simple newspaper applications. For ENMAP we use the following top level elements:

(1) Title: Title of the newspaper
As already indicated
(2) Genre: Newspaper issue
(3) Date of the issue
(4) Language: We use the ISO-639 2b encoding. For each language we use a separate language element with a corresponding font type.
(5) Script or alphabet: Important in case of roman typeface vs. black letter (gothic) fonts. We will use the “script” element of MODS with ISO 15924 codes. Supported values will be “Latn” for “Latin”, “Goth” for “Gothic” and “Cyrl” for “Cyrillic”.
(6) Identifier: EU News Identifier
(7) Identifier: Identifier for descriptive metadata record (e.g. MARC21, ISSN, etc.)

```
<mets:dmdSec ID="MODS_ISSUE_AZ_19260425">
  <mets:mdWrap MDTYPE="MODS">
   <mets:xmlData>
    <mods:mods>
     <mods:recordInfo>
      <mods:recordContentSource>LFT</mods:recordContentSource>
     </mods:recordInfo>
     <mods:genre>newspaper issue</mods:genre>
```

```
      <mods:titleInfo>
        <mods:title>Alpenzeitung</mods:title>
      </mods:titleInfo>
      <mods:identifier type="ENP">LFT_00001</mods:identifier>
      <mods:identifier type="CatalogueIdentifier">486618</mods:identifier>
      <mods:accessCondition>2 - Snippet view</mods:accessCondition>
      <mods:language>
        <mods:languageTerm>German,Italian</mods:languageTerm>
        <mods:scriptTerm>Gothic</mods:scriptTerm>
      </mods:language>
    </mods:mods>
  </mets:xmlData>
 </mets:mdWrap>
</mets:dmdSec>
```

8.2.4 Technical Metadata – NISO MIX

Within ENMAP we use technical metadata only for files that were refined within the project. Also in these cases we use only a small selection of elements which would be possible from the MIX schema.

(1) Compression scheme
(2) Image Width and Height
(3) Optical Resolution
(4) Bit depth
(5) Resolution

```
<mets:techMD ID="AZ_1926_04_25_0001-MASTER-AMD">
  <mets:mdWrap MDTYPE="NISOIMG">
    <mets:xmlData>
      <mix:mix>
        <mix:BasicDigitalObjectInformation>
          <mix:Compression>
            <mix:compressionScheme>COMPRESSION_CCITTFAX4</mix:compressionScheme>
          </mix:Compression>
        </mix:BasicDigitalObjectInformation>
        <mix:BasicImageInformation>
          <mix:BasicImageCharacteristics>
            <mix:imageWidth>3754</mix:imageWidth>
            <mix:imageHeight>5705</mix:imageHeight>
          </mix:BasicImageCharacteristics>
```

```
        </mix:BasicImageInformation>
        <mix:ImageCaptureMetadata>
         <mix:ScannerCapture>
          <mix:maximumOpticalResolution>300</mix:maximumOpticalResolution>
         </mix:ScannerCapture>
        </mix:ImageCaptureMetadata>
        <mix:ImageAssessmentMetadata>
         <mix:ImageColorEncoding>
          <mix:bitsPerSample>
           <mix:bitsPerSampleValue>(1)</mix:bitsPerSampleValue>
           <mix:bitsPerSampleUnit>integer</mix:bitsPerSampleUnit>
          </mix:bitsPerSample>
         </mix:ImageColorEncoding>
        </mix:ImageAssessmentMetadata>
       </mix:mix>
      </mets:xmlData>
     </mets:mdWrap>
    </mets:techMD>
```

8.2.5 Structural Map – SIMPLE

The Structural Map within ENMAP (simple) is very basic and is based on the assumption that the physical order of pages (numbering) is identical with the logical order of pages.

```
<mets:structMap LABEL="Physical Structure" TYPE="PHYSICAL">
   <mets:div ID="phys0" LABEL="Alpenzeitung" TYPE="physSequence"
DMDID="MODS_ISSUE_AZ_19260425">
    <mets:div ID="phys1" ORDER="1" ORDERLABEL="1" TYPE="page">
     <mets:fptr>
      <mets:par>
       <mets:area FILEID="AZ_1926_04_25_0001-MASTER"/>
       <mets:area FILEID="AZ_1926_04_25_0001-ALTO"/>
       <mets:area FILEID="AZ_1926_04_25_0001-VIEWINGCOPY"/>
      </mets:par>
     </mets:fptr>
    </mets:div>
```

8.2.6 FileSection

Naming of FileGroups

(1) FileGroup: ImageGroup

- Subdivisions: OCRMasterFiles, ViewingFiles,

(2) FileGroup: TextGroup

- Subdivisions: ALTOFiles, e.g. NER, TEI,...

Naming of METS files

Files are renamed according to the following schema:

(1) ENP Newspaper ID_ISSUE ID (=Date of issue).

```
<mets:fileSec>
  <mets:fileGrp ID="ImageGroup">
   <mets:fileGrp ID="MasterFiles" USE="Preservation">
    <mets:file ID="AZ_1926_04_25_0001-MASTER" ADMID="AZ_1926_04_25_0001-MASTER-AMD"
     MIMETYPE="image/tiff" SEQ="1" CHECKSUMTYPE="MD5"
     CHECKSUM="DF-4A-90-41-DA-EE-E7-4E-41-4A-E2-F6-7B-0C-AC-EF">
     <mets:FLocat LOCTYPE="URL" xlink:href="file://./OCRmaster/AZ_1926_04_25_0001.tif"/>
    </mets:file>
```

8.2.7 ALTO File

In our case we use the implementation provided by ABBYY FineReader 10. There are several issues with this implementation, but due to the fact that the ALTO Working group plans a new release of ALTO with some updates and changes we will provide a more detailed description in one of the updated versions of this report.

8.3 ENMAP (full)

8.3.1 General considerations

ENMAP (simple) was mainly designed to cope with the task of providing a simple and effective schema for newspaper issues which are OCR processed. But for exploiting the full potential of newspapers a more detailed capturing of information will be important. This is done in the project by CCS for 2 million newspaper pages and by UIBK on the basis of some experiments carried out for some selected newspaper titles.

According to the considerations provided above we use Newspaper Content Items, Newspaper Content Sections and Newspaper Structural Elements as the main concepts for our structural map.

Newspaper content items can not only be regarded as discrete units of the structure. They contain also descriptive metadata which are usually recorded for indexing. In our case this is mostly the title information (title, sub-title, top-title, etc.) as well as author(s), in some cases maybe also the coverage note spatial and the coverage note temporal, or even serial notes for continued novels. Moreover the type of a Newspaper Content Item, such as news, letter to the editor, or job offer, will be available as metadata.

In our design for the structural map we tried to stick to the functional requirements set out in the METS standard. This means that descriptive metadata should go into the descriptive metadata section and structural metadata should be recorded in the structural map.

The following rules can be derived from this general consideration:

1. Newspaper Content Items are represented both in the DMD section of METS as well as in the Structural Map. A news article or a book review are discrete units of content and need therefore be described as “content” units in the MODS section of ENMAP. Often there will be only very limited information available, e.g. only a title and a genre for this NCI, such as “news” or “book review”. In other cases and especially for newspapers of the 20th century more data, such as the author(s), the intro (abstract), a caption line of a photo, and similar information may be available. It is this bibliographic information which is recorded here.

2. Newspaper Structural Elements are represented in the Structural Map of METS. Since Structural Elements are the components of a Newspaper Content Item they are linked to the corresponding areas and regions within the ALTO file. Nevertheless there are some rare cases where this rule needs an exception: Mainly if we think about photos with a caption and an identifiable photographer it makes sense to regard this information as descriptive data and to represent it in the DMD section of the METS container as well.

3. Newspaper Content Sections are somehow in between and the decision to regard them either as being part of the Descriptive Metadata Section of METS or the Structural Map is in our opinion rather arbitrary. There are good arguments to view them mainly as an expression of structural functionality (as providing an additional ordering between Content Items) or as an expression of their descriptive value (e.g. to summarize articles according to the criteria “Foreign affairs”). We have decided to treat them more like Newspaper Content Items and therefore they will also appear in the Descriptive Metadata Section of METS.

The following figure provides an overview:

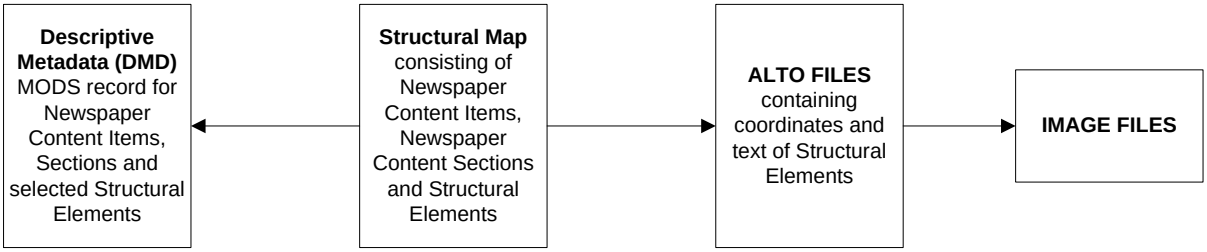


Fig. 1 Structural Map of ENMAP

8.3.2 METS Examples

In the following we provide an example of a structural map built according to the above mentioned concepts: <mets:structMap ID="logical_structure_1" TYPE="logical_structure">

```
<mets:div ID="LS1" TYPE="newspaper_issue" DMDID="MODS_ISSUE_BZN_19100104">
 <mets:div ID="LS2" TYPE="newspaper_content_item" ORDER="1">
  <mets:div ID="LS3" TYPE="title-section" ORDER="1">
   <mets:fptr>
    <mets:area FILEID="FID-BZN_19100104_01-OCRMASTER" COORDS="1 2 3 4"/>
   </mets:fptr>
   <mets:fptr>
    <mets:area FILEID="FID-BZN_19100104_01-ALTO" COORDS="1 2 3 4"/>
   </mets:fptr>
  </mets:div>
 </mets:div>
 <mets:div ID="LS4" TYPE="section">
  <mets:div ID="LS5" TYPE="newspaper_content_item" DMDID="DMD_LS5" ORDER="2">
   <mets:div ID="LS6" TYPE="top_heading" ORDER="1">
    <mets:fptr>
     <mets:area FILEID="FID-BZN_19100104_01-OCRMASTER" COORDS="1 2 3 4"/>
    </mets:fptr>
    <mets:fptr>
     <mets:area FILEID="FID-BZN_19100104_01-ALTO" COORDS="1 2 3 4"/>
    </mets:fptr>
   </mets:div>
   <mets:div ID="LS7" TYPE="heading" ORDER="2">
    <mets:fptr>
     <mets:area FILEID="FID-BZN_19100104_01-OCRMASTER" COORDS="1 2 3 4"/>
```

```
  </mets:fptr>
  <mets:fptr>
   <mets:area FILEID="FID-BZN_19100104_01-ALTO" COORDS="1 2 3 4"/>
  </mets:fptr>
 </mets:div>
 <mets:div ID="LS8" TYPE="heading" ORDER="3">
  <mets:fptr>
   <mets:area FILEID="FID-BZN_19100104_01-OCRMASTER" COORDS="1 2 3 4"/>
  </mets:fptr>
  <mets:fptr>
   <mets:area FILEID="FID-BZN_19100104_01-ALTO" COORDS="1 2 3 4"/>
  </mets:fptr>
 </mets:div>
 <mets:div ID="LS9" TYPE="paragraph" ORDER="4">
  <mets:fptr>
   <mets:area FILEID="FID-BZN_19100104_01-OCRMASTER" COORDS="1 2 3 4"/>
  </mets:fptr>
  <mets:fptr>
   <mets:area FILEID="FID-BZN_19100104_01-ALTO" COORDS="1 2 3 4"/>
  </mets:fptr>
 </mets:div>
 <mets:div ID="LS10" TYPE="paragraph" ORDER="5">
  <mets:fptr>
   <mets:area FILEID="FID-BZN_19100104_02-OCRMASTER" COORDS="1 2 3 4"/>
  </mets:fptr>
  <mets:fptr>
   <mets:area FILEID="FID-BZN_19100104_02-ALTO" COORDS="1 2 3 4"/>
  </mets:fptr>
 </mets:div>
 <mets:div ID="LS11" TYPE="illustration" ORDER="6">
  <mets:fptr>
   <mets:area FILEID="FID-BZN_19100104_02-OCRMASTER" COORDS="1 2 3 4"/>
  </mets:fptr>
  <mets:fptr>
   <mets:area FILEID="FID-BZN_19100104_02-ALTO" COORDS="1 2 3 4"/>
```

```
      </mets:fptr>
    </mets:div>
    <mets:div ID="LS12" TYPE="caption" ORDER="7">
      <mets:fptr>
        <mets:area FILEID="FID-BZN_19100104_02-OCRMASTER" COORDS="1 2 3 4"/>
      </mets:fptr>
      <mets:fptr>
        <mets:area FILEID="FID-BZN_19100104_02-ALTO" COORDS="1 2 3 4"/>
      </mets:fptr>
    </mets:div>
  </mets:div>
```

9. Annex 1 Proposal for a comprehensive list of Newspaper Structural Elements

In the following we provide a list of NSEs together with synonyms, definitions and some remarks concerning their value and the way to capture them.

1. **Newspaper title**

Synonyms

- Name of a newspaper

Definition

- The name of the newspaper as it is indicated on the first page of an issue. The title of a newspaper may change over time and may also be used in several forms. Short name of a newspaper, common acronym of a newspaper, full-name of newspapers.

Value

- For descriptive metadata the value is high, but from the point of view of structural metadata the newspaper title will usually be part of the title section and the running title and therefore be excluded from full-text search or text mining operations.

Caption

- The newspaper title itself is captured during the scanning process, the title section as part of an automated capturing.

2. **Title section**

Definition

- The first part of an issue containing a lot of metadata such as newspaper name, date of publication, imprint information, etc.

Value

- The title section is important for the descriptive metadata that are captured during the scanning process, but from the point of view of structural metadata the title section is just noise for information retrieval and needs therefore to be excluded from full-text search or text mining operations.

Caption

- The title section can be detected automatically since its location and the text itself are repeated in every issue.

3. **Running title**

Synonyms

- Header, column title

Definition

- Comprises typically the first top line of a newspaper page (which spans the columns) and includes often the (short) title of the newspaper, the page number, the issue number, the section heading of the page and the date of publication.

Value

- The value of the running title is rather low since nearly all the information is already available in the metadata which are captured during the scanning process.
- Nevertheless section headings may be interesting for automated retrieval.

Caption

- Running titles are usually automatically located.

4. Date of publication

Definition

- The date when the issue was published.

Value

- The value is high, but usually already known beforehand.

5. Issue number

Definition

- Newspapers are usually numbered either within a year or from the very beginning of their appearance.

Value

- Depending on the policy of a library the issue number may be an important metadata, or it may be just an addition to the date information. Within ENP we are relaying on the publishing date.
- The issue number may appear in the title section as well as in the running title of a page and should therefore be excluded from information retrieval or content based text mining operations.

Caption

- Caption of the issue number is very likely done before the scanning process and/or connected with the recording of the publishing date.

6. Edition statement spatial

Definition

- If newspapers are published at several places than this will be indicated: E.g. London edition, Manchester edition.
- The information on the spatial relation of a newspaper issue.

Value

- This information is a basic requirement for newspapers which have several spatial editions.

Caption

- This information is usually already known at the scanning process and therefore no automated or manual processing will be done.

7. **Edition statement temporal**

Definition

- If newspapers are published several times a day than this will also be indicated: E.g. morning edition, evening edition, etc.

Value

- This information is a basic requirement for newspapers which have several temporal editions per day.

Caption

- This information is usually already known at the scanning process and therefore no automated or manual processing will be done.

8. **Imprint**

Synonyms

- Impressum

Definition

- The imprint contains legal information on the publisher, editor, address, price etc. of the newspaper. It can often be found in the title section of a newspaper.

Value

- Though it might be interesting for historians from an information retrieval point of view it may be seen as noise in the current context.

9. **Section headings**

Synonyms

- Rubric, column title

Definition

- A section heading indicates the general area of content items. It appears either within the running title (column title) or within the page itself.
- Usually it is applied to several content items, but there might also be the case that it appears only for one (in several issues repeated) article.

Value

- As indicated above section headings are important to provide a general subject to the whole page or to a number of articles on a page.
- From an information retrieval point of view they are rather important since their caption will allow to structure results according to the “original subject” of the newspaper.

Caption

- Section headings appear always in several issues and are often stable over many years, e.g. “News from Abroad”, or “Sports”. A list of section headings (and sometimes their order) may be used to automatically detect sections.

10. **Top heading**

Synonyms

- Top 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

Value

- The value is good since the top title explains in more detail the content of an item than the main title.

Caption

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

11. **Heading**

Synonyms

- Head, main title, title

Definition

- Articles within a newspaper usually come with a title indicating the content of the article.
- Larger items may have several titles, such as top heading, sub-heading, inside-heading.

12. **Sub-heading**

Definition

- A title that follows the main title and which provides some additional information.

Value

- The value is high since additional information on the content is provided.

Caption

- Larger content items appear most frequently with a sub-heading which comes in a specific layout and can therefore be automatically captured.

13. **Inside-heading**

Synonyms

- Sub-heading

Definition

- Larger articles are often structured with headings directly within the text body. They are sometimes heads with their own value or sometimes they repeat just a sentence or some words from the body text of the article.

Value

- Sub-heads sometimes appear within sentences and are therefore noise for many applications that relay on tree-taggers and similar tools, sometimes they appear as a regular heading and are therefore additional information.

14. Lead

Synonyms

- Intro, Introduction

Definition

- Usually the first paragraph(s) of a (larger) article providing a summary of the content of the article.

Value

- The value may be rather high for displaying purposes since the lead is a kind of abstract of an item.

Caption

- 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 not be that easy in all cases.

15. Copyright note

Synonyms

- By-line, copyright statement

Definition

- One or more persons responsible for the content of an item. In historical newspapers often authors were not mentioned at all.
- From a historical point of view it is interesting to see that the author information becomes more and more important: It starts with short acronyms for free lancers and photographers and nowadays the full name of the author is usually mentioned for every article within the content section of a newspaper.
- Copyright notes are more often applied to entertainment items, such as novels, poems, cartoons.

Value

- The value is high both on the level of increasing the quality of metadata as well as for information retrieval and post-processing. E.g. NER will benefit from copyright notes but also readers may list all articles of a specific journalist.

Caption

- Author names or abbreviations are important and readers are clearly interested to see “all” articles of a specific journalist within a newspaper.
- Since there are rather strict rules for each newspaper where these copyright notes appear and since the number of authors is limited there is a good chance to find them automatically (“By-line”).

16. **Coverage note spatial**

Synonyms

- Place name

Definition

- Several text types such as news articles commonly indicate the location of the story right at the beginning of the item.

Value

- The value is high since the coverage note indicates for the whole item that there is a relation to a specific location.

Caption

- Due to the fact that coverage notes follow strict rules within one newspaper they can be automatically detected.
- Combined with a NER analysis the quality may even be increased.

17. **Coverage note temporal**

Synonyms

- Date, dateline

Definition

- Some newspapers indicate the exact date of a content item right at the beginning of the item.

Value

- The 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 of high value.

Caption

- Due to the fact that coverage notes follow strict rules within one newspaper they can be automatically detected.
- Combined with a NER analysis the quality may even be increased.

18. **Paragraph**

Definition

- A paragraph is the default unit of a running text and usually provides a single thought or semantic unit.

Value

- The value is high, since a paragraph can be seen as a natural component of a content item.

Caption

- 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.

19. 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.

Value

- The value is high and we know from the history of newspapers that graphical elements are becoming more and more important.

Caption

- Illustrations can be captured with good results in an automated or semi-automated way.

20. 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 displaying stock exchange rates, or TV programmes, etc.

Value

- The value is limited since further processing is often not possible.

Caption

- 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.

21. List

Definition

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

Caption

- The caption may be done automatically.

Value

- The value may be high if the items are taken as named entities and linked to corresponding resources, such as library catalogues.

22. Bibliography

Definition

- Bibliographies are a special kind of a list comprising a number of books or papers.

Caption

- For specific Newspaper Content Items, such as “Rankings of Bestsellers” the caption may be done automatically.

Value

- The value may be high if the items are taken as named entities and linked to corresponding resources, such as library catalogues.

23. Continuation note

Definition

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

Caption

- 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.

Value

- The continuation note itself is noise from the information retrieval point of view. The value of the link between two parts (continuation) is rather high, since it will put together dislocated pieces of an item.

24. Quotations

Definition

- An explicit or implicit record of someone's (verbal) expressions.
- In most cases explicit quotations come with a quotation mark "Yes, we can!", whereas implicit quotations in reported speech may be introduced by typical phrases: He said, that ...

Value

- Whereas explicit quotations are rather seldom in historical newspapers they play an increasing role in modern texts and are of some good value for operations such as text mining.

Caption

- There are several attempts to automatically detect and extract quotations in newspapers.

10. Annex 2: Proposal for a Classification Schema for Newspaper Content Items

10.1 General considerations

Content items follow historical conventions and are very near to “genres” or “text types”. Typical content items are news articles, columns, announcements, letters to the editor, advertisements, weather forecasts, TV programmes, cartoons, photos, etc. There are core items, which can be found in nearly every newspaper, such as “news articles” and “advertisements”, and there are rare items which may occur only in specific newspapers such as aphorisms or jokes.
Newspaper Content Items can be classified according to many criteria. E.g. one could classify content as (1) content provided by professionals and (2) content generated by users. Another classification could be (1) general content vs. (2) paid content.
As already indicated above we will classify content in newspapers according to the inherent message which these articles evoke in a reader. Given this main criterion it makes sense to distinguish five main categories:

- Information
- Opinion
- Entertainment
- Advertisement
- Metadata

It is important to note that we are going for the obvious message of a content item. A newspaper article may be perceived by many readers also as “entertainment” but the main focus is “information”. Such cases may appear in the “Society” section of a newspaper where the actual “news value” is rather low, compared to the entertainment effect.

It is also important to apply historical measures in the classification, not current measures. If e.g. a news article about the political situation at the beginning of the First World War may appear to our (current) mind to be not an “information”, but rather aims at motivating the readers to approve the war than this article still would be classified as “information” and not as “opinion”. It is the historical perspective which needs to be taken, not today’s point of view.

10.2 The five main classes of content items in newspapers

10.2.1 Information

The information category is the core concept within a newspaper. Newspapers were established exactly to provide (latest) news on all aspects of daily live. Therefore news in many forms and formats are summarized under “information”. Since newspapers are published nearly every day the latest news on events or persons are typical for this category.

From the point of view of the user this means that ideally the user shall “know more” or better “understand” the complex reality after having read a news article. In general the ambition is to provide the information in an objective way so that news articles are centred around facts.

10.2.2 Opinion

Opinion items provide a personal expression often based on some information but the facts are structured by the personal viewpoint of the writer. In many cases an opinion item tries to convince or to persuade a reader, or at least it is an (implicit) plea to share the author's opinion. The reader needs to decide: Either to follow this opinion or to reject it. Historically a clear distinction between "information" and "opinion" is rather new and has, for a long time, not been made as explicit as we are expecting this today from a modern newspaper. In historical newspapers we will discover a number of content items which include personal judgement (opinion) without explicitly marking it.

10.2.3 Entertainment

Entertainment items are aiming at the emotional state of a reader and want to change it. Typical entertainment items are serial novels, poems or cross word puzzles. Readers ideally enjoy an item, escape into a fictional sphere, are amused or get involved. Obviously well written news articles and columns are a pleasure for every reader as well, but their entertainment value is usually secondary compared to their general intention – to inform or to convince.

10.2.4 Advertisement

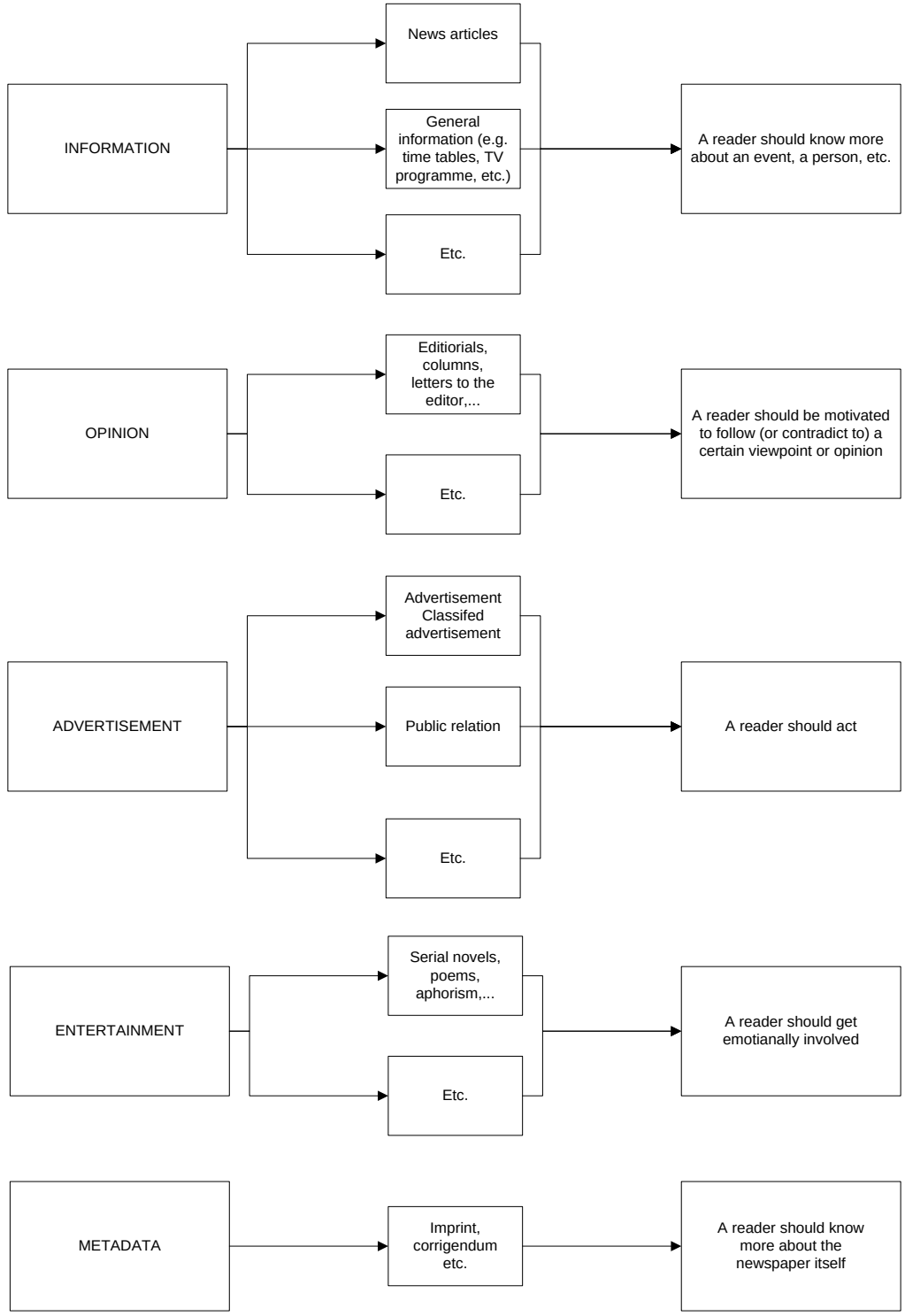
Advertisements typically try to motivate a user to get involved in some kind of action: to buy a certain product, to appraise a certain company, to contact a company with regard to a job offer, to contact a real estate office, to attend a funeral, etc. In the best case the user will actually realize this (more or less implicit) message and act accordingly. Advertisements are paid content and can also be distinguished from edited content via this criterion.

10.2.5 Metadata

Metadata items are self-reflexive and are providing information about the newspaper itself. There are only a few typical content items which are known for this category, e.g. the title section at the top of title pages, the imprint or corrigenda.

The following figure provides a graphical representation of these considerations and lists some examples:

Classification of content items according to their inherent message



Zeichenblatt 1

10.3 Classification schema for Newspaper Content Items

Category	First level	Second level	Definition
1. Information			
	1.1. News		The default class for all Newspaper Content Item in the Information category. Synonyms are “news article”, “article” or “story”.
		1.1.1. Lead Story	The main story, usually on the title page or on the title page of a section.
		1.1.2. Breaking News	The latest news.
		1.1.3. Background News	News that offer a wider view on a certain story, event or person.
		1.1.4. Reportage	Similar to background news, but written from a more personal or subjective perspective.
	1.2. Verbatim reports		
		1.2.1. Interviews	A news article consisting mainly of the verbatim report from an interview (with a well-known person or expert).
		1.2.2. Discussions	A news article consisting mainly of a verbatim report of a discussion of several persons.
		1.2.3. Speeches	A verbatim report from a speech (e.g. laudation or obituary).
	1.3. Biographical news		
		1.3.1. Portraits	A news article about a person.
		1.3.2. Anniversaries	An article on the occasion of an anniversary.
		1.3.3. Obituaries	An article on the occasion of someone’s death. Not to be mixed up with “Death notice” in the “advertisement category”. There it is “paid content”.
	1.4. Factual news		The default category for all news that do not come as a

			narrative, but as lists or tables.
		1.4.1. Weather reports	Reports about the weather and weather forecast.
		1.4.2. Program news	Theatre, music hall, TV, radio programs. Most often organised as a list or table.
		1.4.3. Stock exchange rates	Stock exchange rates
		1.4.4. Railway tables	Timetables of railways, etc.
2. Opinion			
	2.1. Columns		A regular section of a newspaper or magazine devoted to a particular subject or written by a particular person.
	2.2. Editorials		An editorial is a special case of a column and usually written by the chief editor or a well-known author. It often expresses the "official" opinion of the editorial team of a newspaper. Most often it appears at a certain location within the newspaper or on certain days of a week or month, e.g. Saturday.
	2.3. Reviews		A review focuses on a specific literary, artistic or commercial event or product and contains usually a judgement on the value of the reviewed object. The most important reviews in historical newspapers are about books, theatre plays and concerts. Other reviews such as those about fiction books, non-fiction books, audio books, art performances, games, journals, TV plays, theatre plays, radio programs, TV programmes, conferences, concerts, operas, cinema films, etc. are not listed here
		2.3.1. Book review	A review about a (new) book.
		2.3.2. Theatre review	A review about a theatre performance.

		2.3.3. Concert review	A review about a concert.
	2.4. Letters to the editor		Letters to the editor are usually printed in a dedicated section and express the opinion of readers, most often towards an issue raised in the newspaper the days before.
	2.5. Commentaries		Similar to editorials and columns commentaries are most often written by external authors, mostly experts.
3. Entertainment			Entertainment can take many different forms, it might contain literary and artistic works but also cartoons or jokes or cross word puzzles. Since there are so many sub-genres which can be detected in the entertainment section of a newspaper we will not describe them in detail, but just give a short list of the main classes as they appear rather often in (historical) newspapers.
	3.1. Literary works		
		3.1.1. Serial novels	Serial novels are one of the most important classes and can be found in many historical newspapers from the 19th century onwards.
		3.1.2. Poems	A literary wok usually arranged in verses and strophes.
		3.1.3. Theatre plays	Sometimes parts of a theatre play are printed within newspapers.
		3.1.4. Essays	A piece of text mostly with philosophical considerations.
		3.1.5. Aphorism	A short piece of text expressing a sometimes surprising view on a given issue.
	3.2. Graphical works		
		3.2.1. Photos	
		3.2.2. Cartoons	

	3.3. Jokes		A piece of text to cause amusement or laughter, especially a story with a funny punch line.
	3.4. Games		
		3.4.1. Cross word puzzles	A crossword is a word puzzle that normally takes the form of a square or a rectangular grid of white and black shaded squares.
		3.4.2. Riddles	A question or statement intentionally phrased so as to require ingenuity in ascertaining its answer or meaning:
4. Advertisement			
	4.1. Advertisements		A piece of text or a figure used to encourage, persuade, or manipulate readers or to take or continue to take some action.
	4.2. Classified advertisements		Classified advertising is a form of advertising which is particularly common in newspapers. Classified advertisements in a newspaper are typically short, as they are charged for by the line, and one newspaper column wide. The advertisements are grouped into categories or classes such as "for sale—telephones", "wanted—kitchen appliances", and "services—plumbing", hence the term "classified".
		4.2.1. Family notices	Short notices related to families, such as marriage, birth of child, etc.
		4.2.2. Death notices	Short and mostly formal notice on the death of a person.
		4.2.3. Job offers	Announcement of open jobs.
		4.2.4. Real estate offers	Announcements of available real estate objects.
5. Metadata			

	5.1. Imprints		Mostly for legal purposes, a piece of information which informs about the address, the publisher, the editor, etc. of a newspaper.
	5.2. Self-referential notes		A message from the newspaper editor about the newspaper or its content.
	5.3. Billboards		A piece of text attracting the attention of readers in order to read a news article. Billboards are similar to Table of Contents pages in books.