[redacted]

Project Acronym:	Europeana Newspapers
Grant Agreement number:	297380
Project Title:	A Gateway to European Newspapers Online

D5.3 Final public release with updated online resource for documentation

Revision:	1.0
Authors:	**Günther Mühlberger (UIBK)** **Günther Hackl (UIBK)**

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
0.1	23-03-2015	Günther Mühlberger Günther Hackl	UIBK UIBK	Draft
0.2	08-04-2015	Evelien Ket	KB	Internal review
0.3	20-04-2015	Günther Mühlberger Günther Hackl	UIBK UIBK	Following feedback from reviewers
1.0	07-04-2015	Clemens Neudecker, Sandra Kobel	SBB	Internal review and final version

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This deliverable contains original unpublished work except where clearly indicated otherwise.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.

Contents

1. Executive Summary

The following deliverable D5.3 contains the public release of the Europeana Newspaper METS ALTO Profile (ENMAP). Part of the deliverable are also real world examples from several newspaper issues showcasing how the format can be used in enriching historical newspapers with structural metadata. A part of the deliverable is also STRUCTIFY, a tool which can be used to view ENMAP files as well as to generate them.

The focus of ENMAP has been laid to the question of “how to deal with structural metadata” since it turned out during the project that this task is of the highest importance to many libraries which are now facing the challenge to deal with millions of newspaper pages.

As a matter of fact the “deep structuring” of newspapers is just at the very beginning and has not reached a mature status so that it would be too early to actually release a comprehensive standard model. In contrast, we believe that the considerations provided below, together with the examples and a highly flexible software tool are the cornerstones for a standard profile which would cover the complex situation of historical newspapers and be accepted by a large majority of libraries.

2. Introduction

2.1. General purpose

ENMAP was designed to meet two main requirements:

1. To support the ingestion of digitised newspapers so that they can be displayed and searched in a newspaper browsing application. The main challenge herein was that a solution had to be found which fits to all the libraries of the *Europeana Newspaper Project* (ENP). This task was completed with the release of ENMAP (simple) after year 1 of the project. All the data produced in the project are following this format.

2. To provide a comprehensive format that supports a deeper structuring of newspapers. For this purpose, ENMAP advanced was drafted and a first version was released in year 2 of the project[1]. The format was introduced to the public at various information days and events and also discussed in-depth at two workshops organised as part of work package 5.

The current paper is a follow-up of the first release and takes into account the feedback and various suggestions gathered during this phase.

2.2. ENMAP vs. EDM

The *Europeana Newspapers METS ALTO Profile* (ENMAP) which is subject of this report was developed in the *Europeana Newspapers: A Gateway to European Newspapers Online* Project *(ENP)*. Though "Europeana" appears in the title of the format, ENMAP must not be mixed up with the *Europeana Data Model* (EDM).

EDM was designed to support the ingestion of metadata for the *Europeana* portal. The general focus is therefore on descriptive metadata which are collected from a variety of objects such as books, manuscripts, newspapers, images, video- and audio files, and many other types.

In contrast, ENMAP is a format which was developed specifically for digitised newspapers. Also in contrast to EDM, the main objective of ENMAP is to define an *Information Package* which consists of administrative and structural data and includes content files, such as images and text. ENMAP can be seen as an "experimental" format showcasing how

[1] http://www.europeana-newspapers.eu/wp-content/uploads/2014/08/D5.2-Europeana-Newspapers-METS-ALTO-Profile-ENMAP-DRAFT.pdf

Europeana may deal with rich digital objects in the future – but it is not an “official” format as it is the case with the Europeana Data Model.

2.3. ENMAP - simple

Within the *Europeana Newspaper Project* it was necessary to manage the enrichment and delivery of a large set of already digitised newspapers from more than 10 partner libraries. The German company CCS GmbH (CCS) and the University of Innsbruck (UIBK) acted as technical providers and were responsible to enrich 2 million page images (CCS) respectively 8 million page images (UIBK).

In order to set up an effective workflow it was necessary to provide a metadata schema that could support the main objective of the project which was to deliver enriched newspaper pages to a repository and viewing application, the “Newspaper Browser”. This application was developed by *The European Library* (TEL).

The workflow for integrating newspaper files into Europeana respectively into the Newspaper Browser is described in detail in Deliverables 4.5-7[2].

Libraries who are interested to deliver their digitised newspapers to the Newspaper Browser are requested to follow ENMAP simple in order to make the integration as simple and smooth as possible.

2.4. ENMAP - extended

During the course of the project it became obvious that one of the main desiderata with respect to digitised newspapers is less “yet-another” METS profile, but a comprehensive and detailed description of the structural metadata of historical newspapers.

The structural map is the heart of a METS document. It outlines a hierarchical structure for the digital library object, and links the elements of that structure to content files and metadata that pertain to each element[3].

In contrast to books or journals where the Structural Map is more or less straightforward, this is different with newspapers. Newspapers often run over decades or even centuries, they

[2] http://www.europeana-newspapers.eu/public-materials/deliverables/

[3] Cf. Library of Congress. METS An Overview & Tutorial. http://www.loc.gov/standards/mets/METSOverview.v2.html

consist of dozens or even hundreds of single pieces and their layout is often complex and comprises several elements.

A main goal was therefore to find a simple but nevertheless comprehensive description schema which could be applied to any newspaper published in Europe between the 17th and the 20th century. This process is called “deep structuring” and it implies that a “full informational capture”[4] can be carried out. “Full informational capture” means in this context that the information, which is necessary to fully understand the message of the newspaper, can be recorded with the concepts provided by the ENMAP schema.

Due to the fact that a “deep structuring” was not foreseen in the *Europeana Newspapers* project, it was not possible to actually apply the ENMAP extended model to the digitised newspapers of the project except for some sample issues and pages which are part of the online documentation of this deliverable. However, in order to provide more than just the theoretical model, a software tool (STRUCTIFY) has been developed which can be used for several purposes. One of them is to manually generate an ENMAP data structure. Therefore some “real world” example data could be provided which can be used to demonstrate in which way the ENMAP model can be applied to a large number of digitised newspapers.

2.5. Manual and automated enrichment

It is important to understand that the deep structuring of newspapers can hardly be done manually but that it will require support of state-of-the-art methods from computer science. On the other hand the manual intervention will always be the start and end-point of deep structuring. Therefore it was the ambition to provide clear and simple rules which are not only reserved to experts but can also be understood and used by other user groups, such as computer scientists, humanities scholars and volunteers. Their input will be necessary to cope with the immense task of a deep structuring of the European Newspapers collection.

As already indicated, manual structuring serves as the starting point for any project dedicated to deep structuring in two ways:

(1) Layout Analysis, Natural Language Processing and Document Understanding are some of the fields which can be exploited for this task. Independently if expert systems are used or

[4] Stephen Chapman and Anne R. Kenney (1996): Digital Conversion of Research Library Materials. A Case for Full Informational Capture, D-Lib Magazine, October 1996 ISSN 1082-9873 Online: http://www.dlib.org/dlib/october96/cornell/10chapman.html.

if the algorithms and tools are relying on machine learning methods, in both cases a significant number of reference data (Ground Truth) are an essential prerequisite for the development of powerful tools. In other words: If a tool shall be able to detect e.g. "series novels" in historical newspapers it will need a significant number of examples as training and evaluation data. Obviously these training data need to be at least partly generated by hand and shall come in a standardised format, so that both the training data, as well as the output of the tools can be used across the boundaries of a single newspaper or library.

(2) Crowd-sourcing, Citizen Science or simply the involvement of volunteers play a more and more important role in data collection and data reviewing projects. Sites such as Zoomify[5] are allowing users to contribute to the enrichment of data – even for highly complex tasks. Though the library world is still lagging behind these developments – the National Library of Australia with its OCR correction service is one of the few exceptions[6] – it can be expected that this gap will be filled in the next years. Instead of "just" offering OCR text correction it can be argued that in many cases it may be more valuable in terms of improving accessibility to a digitised newspaper that users contribute to the deep structuring of the text. Again it will be necessary to have a common schema available which can be applied to different newspapers.

ENMAP and STRUCTIFY are designed to serve as a common basis which will ease the cooperation between technology providers and libraries, but also between libraries and volunteers by providing a comprehensive metadata schema as well as a tool to generate it.

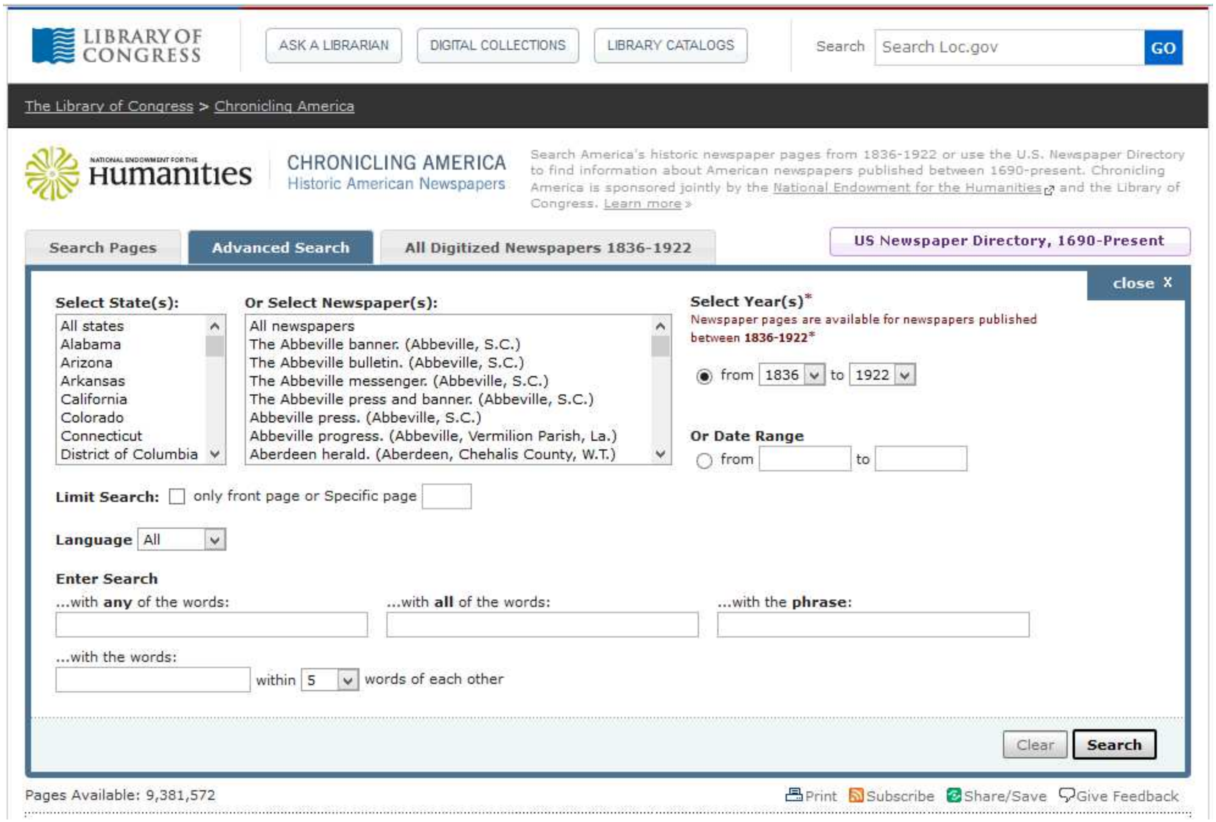
[5] http://www.zoomify.com/
[6] http://trove.nla.gov.au/

3. Use-Cases for ENMAP

3.1. General considerations

Currently there are mainly two ways how digitised newspapers are treated: Either newspapers are scanned, ordered on issue level and enriched with full-text on page level. Or the structuring is done on “article level” which means that all articles are separated and structured. Whereas the first method does not require any manual processing and is therefore rather cost efficient, the second one requires automated processing and manual correction and can therefore be regarded as cost and labour intensive.

The “Chronicling America” project with currently (March 2015) nearly 10 million newspaper pages available online is a perfect example of the first case. The user is able to search in the full-text and may refine his search according to the publication place and publication date. This is actually the same approach as supported by the Europeana Newspaper Browser and the ENMAP simple format.



LIBRARY OF CONGRESS | ASK A LIBRARIAN | DIGITAL COLLECTIONS | LIBRARY CATALOGS | Search | Search Loc.gov | GO

The Library of Congress > Chronicling America

NATIONAL ENDOWMENT FOR THE Humanities

CHRONICLING AMERICA
Historic American Newspapers

Search America's historic newspaper pages from 1836-1922 or use the U.S. Newspaper Directory to find information about American newspapers published between 1690-present. Chronicling America is sponsored jointly by the National Endowment for the Humanities and the Library of Congress. Learn more »

Search Pages | Advanced Search | All Digitized Newspapers 1836-1922 | US Newspaper Directory, 1690-Present

close X

Select State(s):
All states
Alabama
Arizona
Arkansas
California
Colorado
Connecticut
District of Columbia

Or Select Newspaper(s):
All newspapers
The Abbeville banner. (Abbeville, S.C.)
The Abbeville bulletin. (Abbeville, S.C.)
The Abbeville messenger. (Abbeville, S.C.)
The Abbeville press and banner. (Abbeville, S.C.)
Abbeville press. (Abbeville, S.C.)
Abbeville progress. (Abbeville, Vermilion Parish, La.)
Aberdeen herald. (Aberdeen, Chehalis County, W.T.)

Select Year(s)*
Newspaper pages are available for newspapers published between 1836-1922*

● from 1836 to 1922

Or Date Range
○ from ____ to ____

Limit Search: □ only front page or Specific page ____

Language All

Enter Search
...with **any** of the words: ____
...with **all** of the words: ____
...with the **phrase**: ____

...with the words: ____ within 5 words of each other

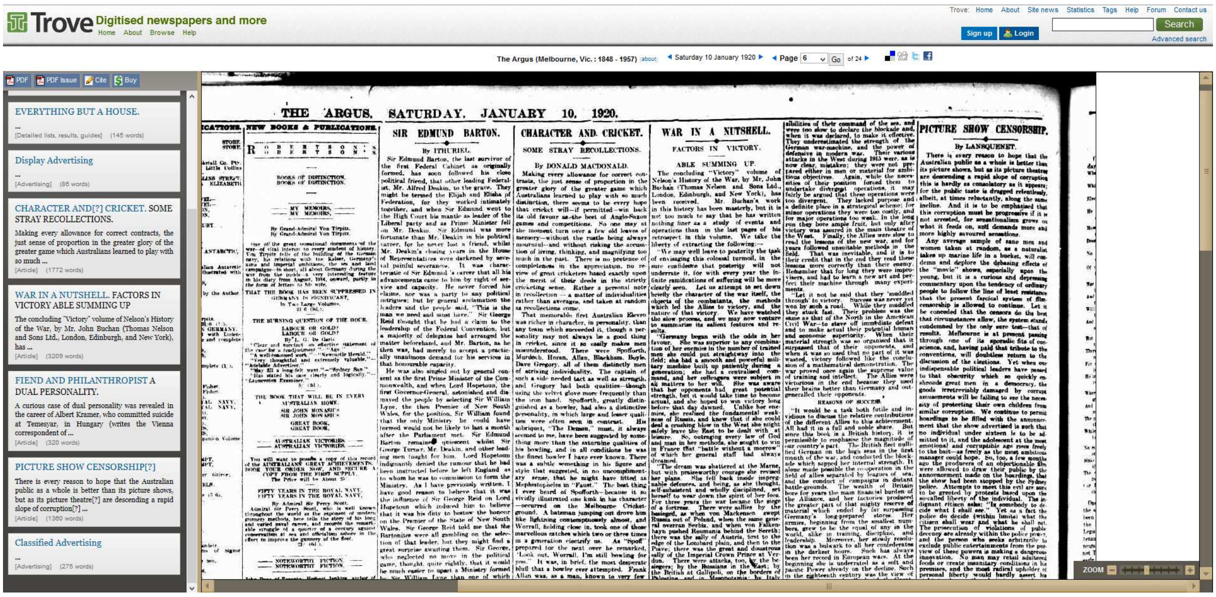
Clear | Search

Pages Available: 9,381,572

Print | Subscribe | Share/Save | Give Feedback

Figure 1: Chronicling America Website

The other pole is represented by the Australian Newspaper project (Trove). Here some structuring is done on “article level” and with some classification: articles, family notices, classified advertisement and similar.



EVERYTHING BUT A HOUSE.
...
[Detailed lists, results, guides] (145 words)

Display Advertising
...
[Advertising] (86 words)

CHARACTER AND[?] CRICKET. SOME STRAY RECOLLECTIONS.
Making every allowance for correct contracts, the just sense of proportion in the greater glory of the greater game which Australians learned to play with so much ...
[Article] (1772 words)

WAR IN A NUTSHELL. FACTORS IN VICTORY ABLE SUMMING UP
The concluding "Victory" volume of Nelson's History of the War, by Mr. John Buchan (Thomas Nelson and Sons Ltd., London, Edinburgh, and New York), has ...
[Article] (3209 words)

FIEND AND PHILANTHROPIST A DUAL PERSONALITY.
A curious case of dual personality was revealed in the career of Albert Kramer, who committed suicide at Temesyar, in Hungary (writes the Vienna correspondent of ...
[Article] (320 words)

PICTURE SHOW CENSORSHIP[?]
There is every reason to hope that the Australian public as a whole is better than its picture shows, but as its picture theatre[?] are descending a rapid slope of corruption[?] ...
[Article] (1360 words)

Classified Advertising
...
[Advertising] (275 words)

Figure 2: Australian National Library: Trove Digitised Newspapers and more

The Australian Newspaper site offers currently more than 16 million digitised pages and claims to make 151 million articles available to the user.

From UIBK’s point of view these two approaches should be complemented with what we call “granularity levels of deep structuring”. This approach would allow handling the structuring in a flexible way so that it can be adapted to the availability of a newspaper search and browsing interface, or the user involvement, the budget of a library and the advances in computer science and technology. The levels which are described below shall demonstrate some of the options for structuring newspapers – it is obvious that there are many other ways one could think of.

By all means deep structuring should be seen as an ongoing process which will come to an end only if the overall goal of “full informational capture” has been reached.

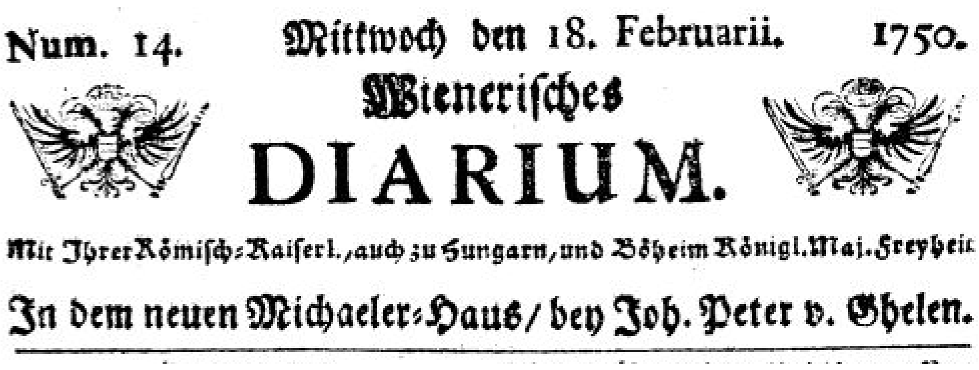
3.2. *Basic enrichment: full-text, physical layout analysis*

This use-case has already been described above. It requires a simple structuring of the newspaper images on a physical level (days, editions) as well as the processing of the images with Optical Character Recognition (OCR). The main challenge in this case is that the basic Layout Analysis as it is part of any OCR engine may be erroneous. E.g. distinct columns might be merged into one, large newspaper titles may be recognised as image instead of text, or the reading order may be confused. The effect on the full-text itself will be marginal, but the effect on more advanced use scenarios will be high as it is shown by the experiments carried out by the University of Salford and which are described in Deliverables D3.5-6[7]. It might therefore be a good idea to invest into an improved layout analysis so that the physical structure of the page is correctly represented as text and image blocks, and maybe also tables and charts.

3.3. Light enrichment: noise reduction, running titles and pictures

Within every newspaper issue some elements can be found which are not directly part of the content but are only included for providing some basic information to the user. These elements are mainly the title section, the running title and the imprint.

Here is an example for an early title section from 1750 containing the issue number, the date, the title, the publication place, the name of the publisher and the rights statement (permission from the emperor):



Num. 14. Mittwoch den 18. Februarii. 1750.

Wieneriſches

DIARIUM.

Mit Ihrer Römiſch-Kaiſerl., auch zu Hungarn, und Böheim Königl. Maj. Freyheit

In dem neuen Michaeler-Haus / bey Joh. Peter v. Ghelen.

Figure 3: Example of a Title section within Wiener Zeitung, 1750

From the point of view of information retrieval all three elements can be regarded as noise. Nearly all of the information such as the title of the newspaper, the editor, the date, the edition, the page number, the issue number, etc. are not only repeated in every issue, but are also recorded in the newspaper catalogue, respectively in the basic structuring of the newspaper on day (and edition) level. To reduce the full-text (and the structural map) from these elements can be regarded as a basic clean up. The following example shows the

[7] http://www.europeana-newspapers.eu/public-materials/deliverables/

negative effect of the title section on the display of the full-text and the structure. Since the title section is very heterogeneous (regarding layout, font type and size) the OCR quality is often very bad. Unfortunately it is in many cases the first impression of a user who looks at the issue of a historical newspaper. The full-text of the example above starts in the following way:

> *Kum. 14. Mittwoch den 18. kebrusrii«. 1750.*
> *Mit?hrerR§misch-RaiserI.,auchzuHungar«,»nd BZHe!mRömgl.Maj.Freyheic*
> *Zn dem neuen Michaeler-HauS/ bey Zoh. Peter v. Ghelen.*

It has to be taken into account that this effect can be seen at about 3.5 million enriched issues within the Europeana Newspaper project.

From the point of view of pattern recognition one can imagine that it is a rather trivial task to detect such title sections, either based on specific rules (e.g. title sections always appear at the top of the first page and are repeated every issue with only minor changes) or by machine learning (e.g. one could tag some hundreds of title sections and use them as input for a learning algorithm).

Similar observations can be made for running titles which may contain some extra information, such as a “section heading” or “rubric”. A closer look on this subject will be provided at the next granularity level.

Though the “imprint” is in most cases hidden within the running text of an issue it consists for many years of the same or very similar piece of text which again can be used as input for a pattern recognition system which utilises not only features based on layout information, but also on textual information.

With the improvement of printing technologies, newspapers started to include photographs during the first decades of the 20th centuries. The basic layout analysis as it is included in OCR engines distinguishes automatically between text and pictures. This information, which is usually stored in the ALTO file (or a similar XML output from the OCR engine), can easily be utilised to enrich digitised newspapers with this kind of information. This will improve browsing capabilities (“Show me all photographs of this issue”), but also search features (“Show me all pictures where a specific keyword can be found “nearby””), i.e. some dozens of words above and below to the picture. The following screenshot shows the implementation of such a search query which is also rather similar to the image search of Google:

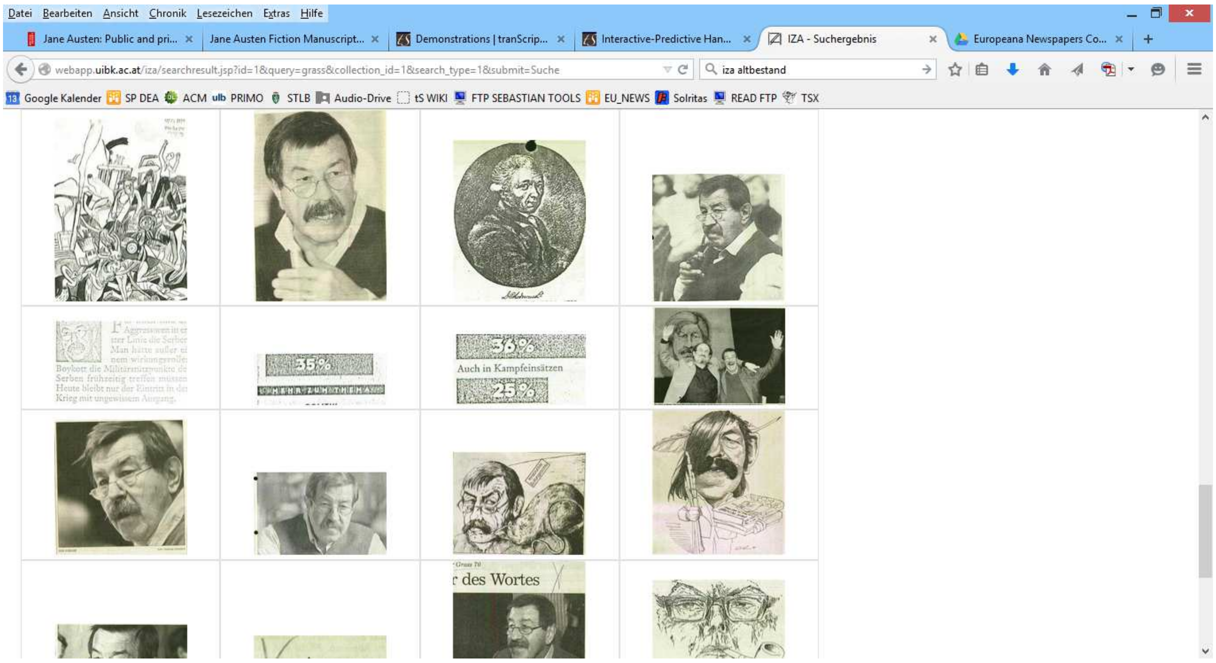
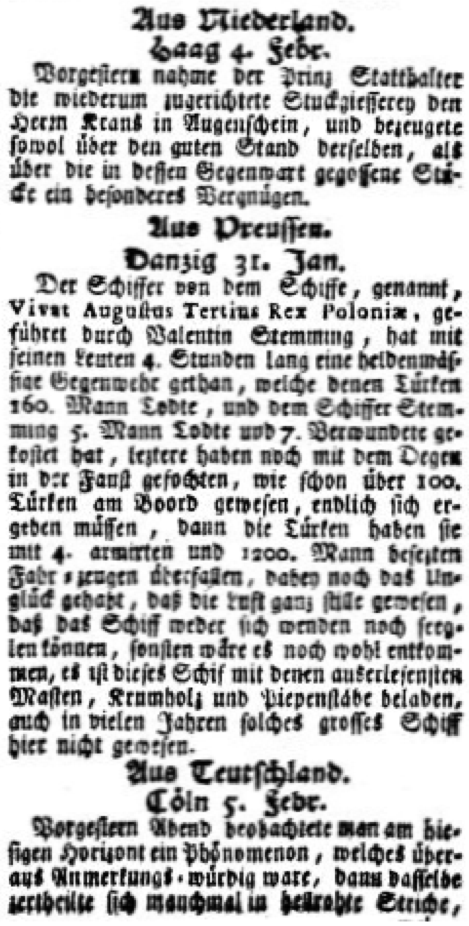

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
Jane Austen: Public and pri...
Jane Austen Fiction Manuscript...
Demonstrations | tranScrip...
Interactive-Predictive Han...
IZA - Suchergebnis
Europeana Newspapers Co...
webapp.uibk.ac.at/iza/searchresult.jsp?id=1&query=grass&collection_id=1&search_type=1&submit=Suche
iza altbestand
Google Kalender
SP DEA
ACM
ulb PRIMO
STLB
Audio-Drive
tS WIKI
FTP SEBASTIAN TOOLS
EU_NEWS
Solritas
READ FTP
TSX
35%
36%
Auch in Kampfeinsätzen
25%
des Wortes

Figure 4: Example of a full-text search taking into account the illustrations of newspapers

The overall effect of this noise reduction, cleaning and simple enrichment may not be tremendous but one has to see that it will have a positive effect on information retrieval and user friendliness especially compared with the limited effort which is required to realise this granularity level.

3.4. *Advanced enrichment: sections*

If we look at historical newspapers we can easily see that their main principle of structuring the content was to indicate from where the news were coming from and therefore to list the news according to their place of origin. E.g. the page below is taken again from 1750 from the “Wiener Zeitung”:



Aus Niederland.

Haag 4. Febr.

Vorgeſtern nahme der Prinz Statthalter die wiederum zugerichtete Stuckgieſſerey den Herrn Krans in Augenſchein, und bezeugete ſowol über den guten Stand derſelben, als über die in deſſen Gegenwart gegoſſene Stücke ein beſonderes Vergnügen.

Aus Preuſſen.

Danzig 31. Jan.

Der Schiffer von dem Schiffe, genannt, Vivat Augustus Tertius Rex Poloniæ, geführet durch Valentin Stemming, hat mit ſeinen Leuten 4. Stunden lang eine heldenmäſſige Gegenwehr gethan, welche denen Türken 160. Mann Todte, und dem Schiffer Stemming 5. Mann Todte und 7. Verwundete gekoſtet hat, leztere haben noch mit dem Degen in der Fauſt gefochten, wie ſchon über 100. Türken am Boord geweſen, endlich ſich ergeben müſſen, dann die Türken haben ſie mit 4. armirten und 1200. Mann beſezten Fahr-zeugen überfallen, dabey noch das Unglück gehabt, daß die Luft ganz ſtille geweſen, daß das Schiff weder ſich wenden noch ſeeglen können, ſonſten wäre es noch wohl entkommen, es iſt dieſes Schif mit denen auserleſenſten Maſten, Krumholz und Piepenſtäbe beladen, auch in vielen Jahren ſolches groſſes Schiff hier nicht geweſen.

Aus Teutſchland.

Cöln 5. Febr.

Vorgeſtern Abend beobachtete man am hieſigen Horizont ein Phönomenon, welches überaus Anmerkungs-würdig ware, dann daſſelbe zertheilte ſich manchmal in hellrohte Striche,

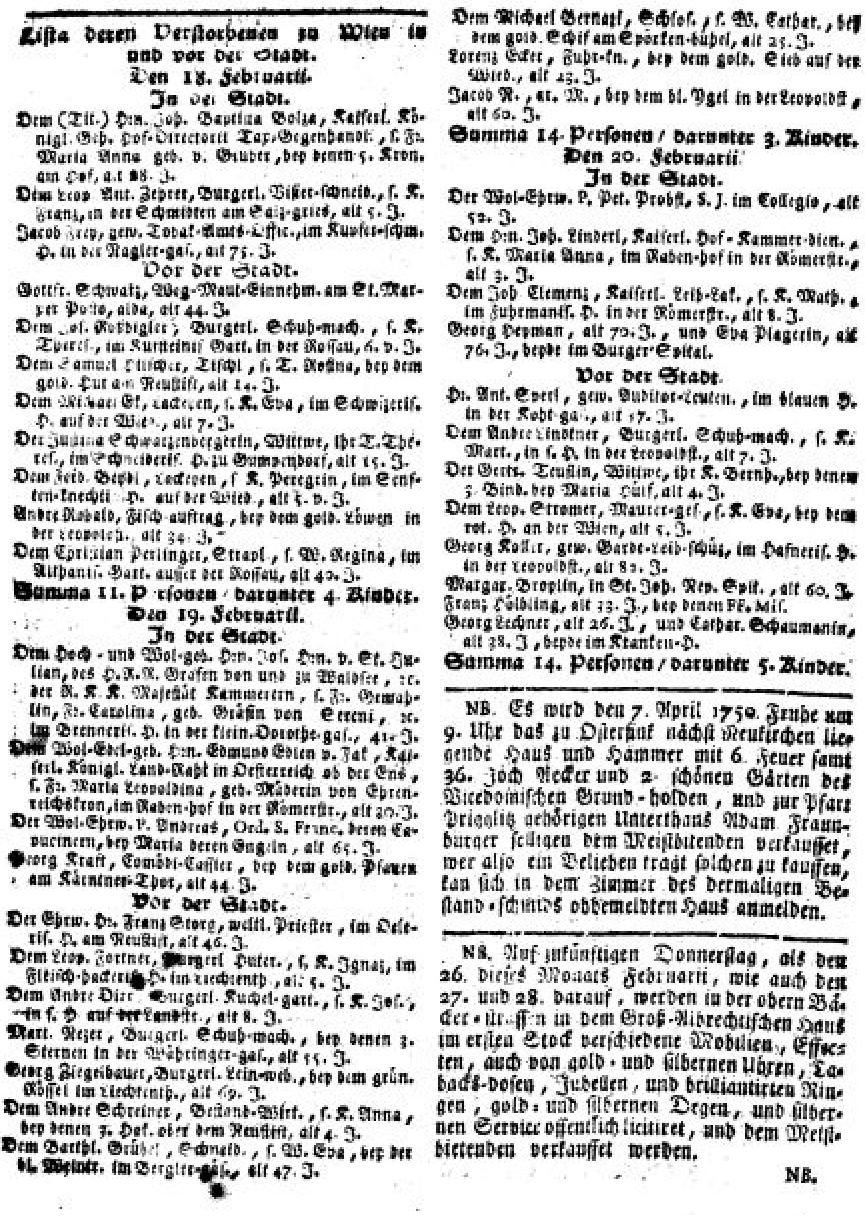
Figure 5: Wiener Zeitung, 1750

The news are ordered according to: “News from the Netherlands, Prussia and Germany”.

So the same headings appear over decades and it took until the 19th century before single content pieces were also indicated with a specific title and more information was added, such as sub-titles, or the detailed name of the information source. The structuring of newspapers into sections is therefore one of the basic layout principles of every newspaper and is still utilised until today.

From a technical point of view these sections can be exploited for automated structuring: Since the section headings remain the same and appear often in every issue (“Aus Teutschland”, …) it will be possible to detect them in an automated way, even if the amount of OCR errors may be significant.

Moreover, from the very beginning of newspaper publishing, specific sections are reserved for content which may not be seen as “classical news”, e.g. official announcements and classified advertisements. Examples are “Lists of decedents” or “New books arrived” – all these news are also put into sections and are repeated over decades and centuries. Again an early example from 21 February 1750 of a “List of persons who died in Vienna”:



Liſta deren Verſtorbenen zu Wien in und vor der Stadt.

Den 18. Februarii.

In der Stadt.

Dem (Tit.) Hrn. Joh. Baptiſta Bolza, Kaiſerl. Königl. Geh. Hof-Directorii Tax-Gegenhandl., ſ. Fr. Maria Anna geb. v. Gruber, bey denen 5. Kron. am Hof, alt 88. J.
Dem Leop. Ant. Zepter, Burgerl. Biſier-ſchneid., ſ. K. Franz, in der Schmidten am Salz-gries, alt 5. J.
Jacob Krep, gew. Tobak-Amts-Offic., im Kupfer-ſchm. H. in der Nagler-gaſ., alt 75. J.

Vor der Stadt.

Gottfr. Schwatz, Weg-Maut-Einnehm. am St. Marxer Poſto, alda, alt 44. J.
Dem Joſ. Rosbigler, Burgerl. Schuh-mach., ſ. K. Thereſ., im Kurtzeiniſ. Gart. in der Roſſau, 6. v. J.
Dem Samuel Fiſcher, Tiſchl., ſ. T. Roſina, bey dem gold. Hut am Neuſtift, alt 14. J.
Dem Michael Ek, Lackeyen, ſ. K. Eva, im Schweizeriſ. H. auf der Wied., alt 7. J.
Der Juliana Schwarzenbergerin, Wittwe, ihr T. Thereſ., im Schneideriſ. H. zu Gumpendorf, alt 15. J.
Dem Ferd. Geppl, Lackeyen, ſ. K. Peregrin, im Senften-knechtiſ. H. auf der Wied., alt 5. v. J.
Andre Rabald, Fiſch-auftrag., bey dem gold. Löwen in der Leopoldſt., alt 34. J.
Dem Chriſtian Perlinger, Straßl., ſ. W. Regina, im Althaniſ. Gart. auſſer der Roſſau, alt 40. J.

Summa 11. Perſonen / darunter 4. Kinder.

Den 19. Februarii.

In der Stadt.

Dem Hoch- und Wol-geb. Hrn. Joſ. Hrn. v. St. Julian, des H. R. R. Grafen von und zu Waldſee, rc. der R. K. K. Majeſtät Kammerern, ſ. Fr. Gemahlin, Fr. Carolina, geb. Gräfin von Cereni, rc. im Brenneriſ. H. in der klein. Dorothe-gaſ., 41. J.
Dem Wol-Edel-geb. Hrn. Edmund Edlen v. Jak, Kaiſerl. Königl. Land-Rath in Oeſterreich ob der Ens, ſ. Fr. Maria Leopoldina, geb. Maderin von Ehrenreichskron, im Raben-hof in der Römerſtr., alt 30. J.
Der Wol-Ehrw. P. Andreas, Ord. S. Franc. deren Capucinern, bey Maria deren Englein, alt 65. J.
Georg Kraft, Comödi-Caſſier, bey dem gold. Pfauen am Kärntner-Thor, alt 44. J.

Vor der Stadt.

Der Ehrw. Hr. Franz Storg, weltl. Prieſter, im Oeſteriſ. H. am Neuſtift, alt 46. J.
Dem Leop. Fortner, Burgerl. Huter., ſ. K. Ignaz, im Fleiſch-hackeriſ. H. im Liechtenth., alt 5. J.
Dem Andre Dier, Burgerl. Kuchel-gart., ſ. K. Joſ., im ſ. H. auf der Landſtr., alt 8. J.
Mart. Rezer, Burgerl. Schuh-mach., bey denen 3. Sternen in der Währinger-gaſ., alt 55. J.
Georg Ziegelbauer, Burgerl. Lein-web., bey dem grün. Röſſel im Liechtenth., alt 69. J.
Dem Andre Schreiner, Beſtand-Wirt., ſ. K. Anna, bey denen 3. Hof. ober dem Neuſtift, alt 4. J.
Dem Barthl. Grüber, Schneid., ſ. W. Eva, bey der bl. Weintr. im Bergler-gaſ., alt 47. J.
Dem Michael Bernatzl, Schloſ., ſ. W. Cathar., bey dem gold. Schif am Spörken-bühel, alt 25. J.
Lorenz Ecker, Fuhr-kn., bey dem gold. Stieb auf der Wied., alt 43. J.
Jacob R., ar. M., bey dem bl. Vgel in der Leopoldſt., alt 60. J.

Summa 14. Perſonen / darunter 3. Kinder.

Den 20. Februarii.

In der Stadt.

Der Wol-Ehrw. P. Pet. Probſt, S. J. im Collegio, alt 52. J.
Dem Hrn. Joh. Linderl, Kaiſerl. Hof-Kammer-dien., ſ. K. Maria Anna, im Raben-hof in der Römerſtr., alt 3. J.
Dem Joh Clemenz, Kaiſerl. Leib-Lak., ſ. K. Math., im Fuhrmaniſ. H. in der Römerſtr., alt 8. J.
Georg Heyman, alt 70. J., und Eva Plagerin, alt 76. J., beyde im Burger-Spital.

Vor der Stadt.

Hr. Ant. Eperl, gew. Auditor-Leuten., im blauen H. in der Kohl-gaſ., alt 57. J.
Dem Andre Lindner, Burgerl. Schuh-mach., ſ. K. Mart., in ſ. H. in der Leopoldſt., alt 7. J.
Der Gertr. Teuflin, Wittwe, ihr K. Bernh., bey denen 3. Bind. bey Maria Hülf, alt 4. J.
Dem Leop. Stromer, Maurer-gſ., ſ. K. Eva, bey dem rot. H. an der Wien, alt 5. J.
Georg Koller, gew. Garde-Leib-ſchüz, im Hafneriſ. H. in der Leopoldſt., alt 82. J.
Margar. Broplin, in St. Joh. Nep. Spit., alt 60. J.
Franz Hölbling, alt 33. J., bey denen PP. Miſ.
Georg Lechner, alt 26. J., und Cathar. Schaumanin, alt 38. J., beyde im Kranken-H.

Summa 14. Perſonen / darunter 5. Kinder.

NB. Es wird den 7. April 1750. Fruhe um 9. Uhr das zu Oſterhuk nächſt Neukirchen liegende Haus und Hammer mit 6. Feuer ſamt 36. Joch Aecker und 2. ſchönen Gärten des Vicedomiſchen Grund-holden, und zur Pfarr Prigglitz gehörigen Unterthans Adam Fraunburger ſelligen dem Meiſtbitenden verkauffet, wer alſo ein Belieben tragt ſolchen zu kauffen, kan ſich in dem Zimmer des dermaligen Beſtand-ſchmids obbemeldten Haus anmelden.

NB. Auf zukünftigen Donnerſtag, als den 26. dieſes Monats Februarii, wie auch den 27. und 28. darauf, werden in der obern Bäcker-ſtraſſen in dem Groß-Albrechtiſchen Haus im erſten Stock verſchiedene Mobilien, Effecten, auch von gold- und ſilbernen Uhren, Tabacks-doſen, Jubellen, und brilliantirten Ringen, gold- und ſilbernen Degen, und ſilbernen Service offentlich licitiret, und dem Meiſtbietenden verkauffet werden.

NB.

Figure 6: Wiener Zeitung, 1750, List of people who died recently

If successful, the benefit of this structuring process for many applications will be high: Users may be able to access specific content stemming from “London”, “Paris”, or “Vienna”, or content from a “classified advertisement” section. The place names may be matched with geolocations so that a user may see at first glance from where the majority of news stems within a given newspaper.

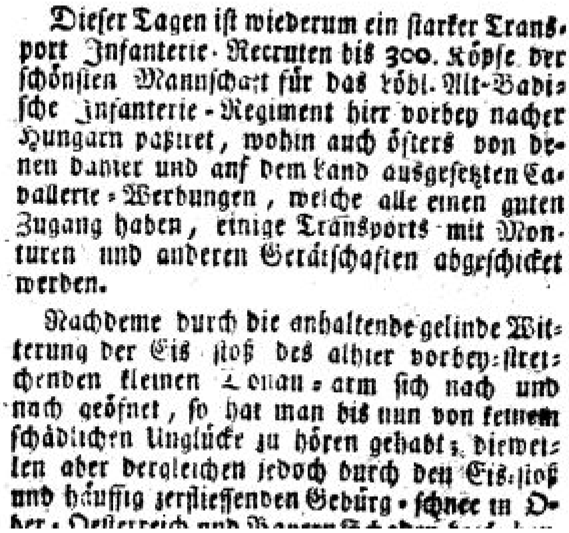
Other benefits may be realised via a facetted search where the search hits are ordered also according to their appearance in a specific section, e.g. “Show me all occurrences of my search query in ‘News from abroad’”.

Also it is to expect that users will have a high interest in customisable export functions: “Export all text from 1789 subsumed under the heading ‘Paris’ or ‘France’”. It can be assumed that especially humanities scholars will benefit from this kind of enrichment but also be willing to contribute to this task – since it can easily be explained and demonstrated.

3.5. Full enrichment: articles and detailed metadata

Apart from a rough structuring into sections, newspapers contain a large number of single information pieces. In early years these single pieces are often just separated by paragraphs, and are therefore hard to discriminate, whereas in later years separators and other layout features (bold, brackets, letter spacing, etc.) are used to indicate the start of a new piece.

An example from the issue above:



Dieſer Tagen iſt wiederum ein ſtarker Transport Infanterie-Recruten bis 300. Köpfe der ſchönſten Mannſchaft für das löbl. Alt-Badiſche Infanterie-Regiment hier vorbey nacher Hungarn paßiret, wohin auch öfters von denen dahier und auf dem Land ausgeſetzten Cavallerie-Werbungen, welche alle einen guten Zugang haben, einige Transports mit Monturen und anderen Geräthſchaften abgeſchicket werden.

Nachdeme durch die anhaltende gelinde Witterung der Eis ſtoß des alhier vorbey-ſtreichenden kleinen Donau-arm ſich nach und nach geöfnet, ſo hat man bis nun von keinem ſchädlichen Unglücke zu hören gehabt; dieweilen aber dergleichen jedoch durch den Eis-ſtoß und häuffig zerflieſſenden Gebürg-ſchnee in Ober-Oeſterreich und Bayern

Figure 7: Articles separated by paragraphs

The paragraph marks two completely different news items, such as the report about a regiment and the fact that the Danube had frozen and had therefore caused problems with shipping.

With the further development of newspapers in the late 19th century, individual articles are getting more important – reflecting the fact that newspapers offer more and more information so that readers start to select the content, instead of reading an issue from the first to the last page. In order to support users in navigating through a newspaper page, individual articles are now indicated with a heading to provide the first information about the content of the news.

A rough estimation shows that the number of distinguishable content pieces within one newspaper issue increases dramatically over the years. In a newspaper of the 18th century some dozens of articles and advertisements can be found, already in the 19th century hundreds of single news can be found and at the heyday of newspapers in the 20th century even some thousand news are included in one issue. This development goes along with an enlargement of the paper size and a reduction of the font size so that more and more text can be delivered to the reader.

It is obvious that a correct separation of a digitised newspaper not only into sections, but also into individual news articles and similar contributions is the final objective of the advanced enrichment task.

4. Concepts

4.1. Introduction

The naming of the main concepts was a sophisticated process and led to a number of discussions. The main task which needs to be solved is the following: When talking about newspapers the usual concept which comes to mind is the “(news) article”: A distinct piece of content which is new and more or less independent from other content in this specific issue of a newspaper.

On the other hand, if one looks at a newspaper as a whole, there are many content pieces which can hardly be classified as an “article”, e.g. “lists of ships arrived”, or a “job offer”, or the “weather report” or a “commercial advertisement”. These pieces which are usually classified as advertisements - or in German “Anzeige” (“announcement”) - are clearly separate from “articles” and would be excluded when sticking to “article” as the overall term. UIBK tried to find a term which subsumes both names and which emphasises – from a more abstract point of view – the concept behind articles and classified advertisements.

The suggestion is to use the term “Content Unit” in order to express that a newspaper issue can be seen as a collection of distinct pieces which are separated by each other by their content. A narrative about event A can be separated from a narrative about an event B, but it is on the same level as an offer for job A and another offer for job B.

A similar problem appeared when dealing with the layout of a newspaper. Obviously there are some parts which have a very distinct layout, such as a large title or a sub-title of an article, a lead, a caption, a by-line, and several other elements. Also in this case the task was to find a term which is able to describe all heterogeneous elements. In the first version of ENMAP the expression “Structural elements” was used, but later changed – mainly to express the relationship between the concepts in the current document - into “Content Item”. Such “Content Items” are parts or elements of “Content Units”.

Finally the sections within newspapers had to be named, but in this case it was also possible to keep the usual term and call it therefore “Content Sections”.

If UIBK speaks about Content Units and Content Sections in a broader way, the term “content pieces” is used. If UIBK speaks about all structural (and partly descriptive) metadata, the term “structural features” is used in the broadest way. The process itself is named “deep structuring”.

In short: UIBK is fully aware that – apart from Content Sections which is a common term in newspapers – the suggestions to introduce two new terms, Content Units and Content Items may lead to some discussions. On the other hand “new terms” have the advantage that they are in some way unburdened from everyday language and therefore the temptation to use them in a different way than described here may be lower.

4.2. Content Units

A Newspaper Content Unit is a distinct piece of content within a newspaper often in the form of a narrative, but also appearing in other forms, such as an announcement or advertisement. 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, the commercial advertisements, or in the letters to the editor section. Each single job announcement and each single letter is one Content Unit. The main criteria to make a distinction between two Content Units is the content, may it be a long list of stock exchange rates, or job announcements or “ships arrived today”. A paragraph in an article, or a column in the stock exchange table are pieces of the Content Unit, but are not messages on their own – they need some context which is provided by the “rest” of the article to be fully understood.

Content Units are most often also intellectual entities in the sense that the “copyright” or the “editorial responsibility” can be clearly specified and allocated to the editorial team which may be a person or an organisation (news agency, other newspaper). In the 20th century contributors such as authors/journalists, photographers, illustrators or cartoonists, are explicitly marked in the article whereas historical newspapers do very rarely mention the actual writer.

In many cases the layout of a newspaper indicates the “borders” between Content Units. E.g. separators (lines or bullet points) are used between articles or the headline indicates the start of a “new” unit. Nevertheless the layout is only one criterion among others to classify a Content Unit, the main criteria is the content.

4.3. Content Sections

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, others only on Saturday. The fact that they are repeated is the most important distinction to Newspaper Content Units, which are per se unique. Though every newspaper developed its own “vocabulary”, similar sections appear at several newspapers at a time. E.g. one can find “Local news” vs. “News from abroad”, “Death notices” and “New books” and similar sections in nearly every newspaper at a given period of time.

A Newspaper Content Section can also be seen as a collection of several Content Units. 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 Content Units, sections neither have a distinct message, nor are they an intellectual unit. They are a compilation of single messages where these messages are put together by some (rather arbitrary) criteria (“Latest telegrams”). Newspaper Content Sections may be better compared to the functionality of a “subject heading”, or an “indexing term” that specifies an aspect many content units share.

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

4.4. Content Items

Newspaper Content Items are the third main concept of the scheme and are the single pieces which are used to build a Content Section or Content Unit. Examples are headlines, sub-headlines, leads, pictures, copyright notes, paragraphs, tables, etc.

Newspaper Content Items 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 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 role of Content Items is therefore a kind of “meta-message” for the reader. It aims mainly at supporting him or her in understanding the content and being able to navigate through the complex layout of a newspaper. Due to the fact that the repertoire of structural elements was developed over a long period of time, a specific semantics connected to the layout is associated with Content Items. E.g. even when looking at a newspaper from far away, or in a completely foreign language, the semantics of some Content Items will be understood, such as headlines, sub-titles, caption lines, etc. even without understanding a single word. It is exactly this aspect that makes structural elements so interesting for automated processing and enhancement via Optical Layout Recognition.

Content Items 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 Unit or a Newspaper Content Section.

In short: If someone were to 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 target media. The Newspaper Content Unit would be part of the book edition, not the actual representation or manifestation of a given article.

4.5. Hierarchical Structuring

A newspaper issue may consist of several sections which include several other sections and units. Not in all cases will it be easily possible to draw a clear line between the hierarchies.

In order to make the usage of the concepts as easy as possible, the following rules are suggested:

(1) The basic section of a newspaper is the issue itself. Issues are repeated over a longer period and with a similar structure.

(2) Content Sections may incorporate other Content Sections and Content Units. The same is true for Content Units, i.e. a large unit (article) may also contain other units.

(3) Content Sections and Content Units need not to be hierarchically nested, i.e. it is not necessary that every unit is part of a section (apart from the issue itself).

(4) Sections and units need to contain at least one Content Item, i.e. a paragraph, or text region.

5. General Classification of content within newspapers

5.1. Background

Another important aspect of the deep structuring is that one can easily observe that the content which is included in a newspaper can be rather easily classified according to several criteria.

Looking at the historical development of newspapers, an impressive differentiation process can be observed: Starting with newspapers from the 17th century which comprise just 4 or 6 pages with content that is placed at one column and structured with a few section headings, as seen already in the 18th century newspapers with 12 or more pages, with two columns, and some specific sections containing job offers, lists of published books and other classified advertisements.

In the 19th century newspapers not only the number of pages increases significantly but several new elements appear, e.g. pages are structured into several columns, the usage of sections, articles, advertisements, classified advertisements, becomes more sophisticated and complex. Finally, in the first half of the 20th century photographs are included and the layout of newspapers finds its modern form, with large headings, sophisticated layout and – especially at Saturdays – impressive amounts of pages[8].

If the content is classified within this historical development one can mainly distinguish five classes: Information (news), advertisements (including classified advertisements), entertainment, opinion and in order to complete this list, the "metadata" class also has to be mentioned, i.e. all the content pieces which are just dedicated to inform the user about the content of the newspaper itself, but also about the publisher, the date of publication and the price of the issue.

The main criteria which are used for this simple classification scheme is the inherent meta-message of the content which is directed towards the reader: In the case of "information" the user will get some news, in the case of "entertainment" the user will be emotionally affected, in the case of "opinion" the user may decide to follow or to reject the message, in the case of "metadata" the user gets some information about the newspaper itself.

[8] Cf. http://lab.softwarestudies.com/2012/03/visualizing-newspapers-history-hawaiian.html

It is important to note that UIBK is interested in the obvious message of a content item. A newspaper article may be perceived by many readers 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.

Moreover, UIBK used a historical perspective for applying this 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 the (current) mind not to be “information”, but rather aiming at motivating the readers to approve the war, then this article still would be classified as “information” and not as “opinion”. It is the historical perspective which shall be reconstructed, not today’s point of view.

5.2. Information

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

For the self-conception of the newspaper as being the fourth pillar in a democratic society, the criteria to provide independent, reliable, well-investigated and truly unique news became fundamental and still is the dominant attitude – though with the advent of the internet, new independent publishing (blogs, social media) and the economic crisis of printed newspapers, this concept may need some revision.

5.3. Advertisement

Together with the first category, advertisements are from a historical perspective the second important building block for the content of newspapers. In contrast to information, where the news are created and edited by the editorial team, advertisements are external content and usually redacted by a person or institution which is not directly responsible for the editing process of a newspaper, e.g. the printer.

Advertisements play a substantial role for the business model since their content is usually paid content and one of the main sources of income for the newspaper publisher – another clear distinction to the edited content which forms the “information” category.

Official announcements from administrative or political bodies, commercial advertisements and classified advertisements appear from the beginning of newspaper publishing and are also one criterion to distinguish newspapers from regular commercial or diplomatic reports which are also providing news on similar aspects.

Looking at the inherent message of advertisements, they typically try to motivate a user to get involved in some kind of action: to behave in a certain way (official announcements and declarations), to buy a certain product, to contact a company with regard to a job offer, to contact the owner of a flat, to attend a funeral, etc. In the best case the user will actually realise this (more or less implicit) message and act accordingly.

5.4. Entertainment

Whereas information and advertisements are the core elements of a newspaper, entertainment only appears in the mid of the 19th century. Nevertheless, for some literary genres newspapers became *the* publishing platform: E.g. writers such as Victor Hugo, Charles Dickens, Gustave Flaubert or Leo Tolstoy published many of their novels first in newspapers. A similar effect can be observed with cartoons which became popular at the beginning of the 20th century.

Entertainment can be easily separated from “Information” and “Advertisements” for several reasons. Again it is external content, not produced by the editorial team. Moreover, entertainment aims to capture the attention of readers by addressing their emotion. Typical entertainment pieces are serial novels, poems or cross word puzzles. Readers ideally enjoy a poem, escape into a fictional sphere with a series novel, or are amused with a joke or get involved with a cross-word puzzle. 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.

5.5. Opinion

Also content which is directly marked as the point of view of either the editor, or a well-known person, is an innovation which appears in newspapers in the second half of the 19th century. The feuilleton is the classical example. It occupied its place at the first page of a newspaper and was clearly separated from other content. Other early examples of clearly marked opinions are book and theatre reviews, which appear frequently in the second half of the 19th century.

Opinions can be characterised as a personal expression often based on some information, but the facts are structured by the subjective viewpoint of the writer. In many cases a piece of opinion 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 a lot of content can be discovered which includes personal judgement (opinion) without explicitly marking it – but this will be subject to some opinion mining rather than deep structuring according to the overall goal of reconstructing a "full informational capture".

5.6. Metadata

To complete the list, content which can be classified as "metadata" also has to be mentioned. This is motivated by the fact that published content always needed a minimum of "responsibility declaration". In order to be able to publish a newspaper, the permission of the government or some legal requirements had to be fulfilled – such as who is in charge of the newspaper, where is the place of publication, the address, the price and similar information. This type of meta-information appears usually in the title section at the first page of an issue, but also in an imprint section which can be found within the issue.

A perfect example for these considerations can be seen in an edition of the Innsbrucker Nachrichten, from the 4th of June 1870.

In the figure below four of five main classes are used: Metadata (in blue) containing some information on the newspaper itself within the Title section, Information (in yellow), with the edited content (news), the advertisement section (in red) comprising not only commercial advertisements, but also classified advertisements and official announcements (edicts, auctions) and finally the Entertainment section (in green) with a classical series novel.

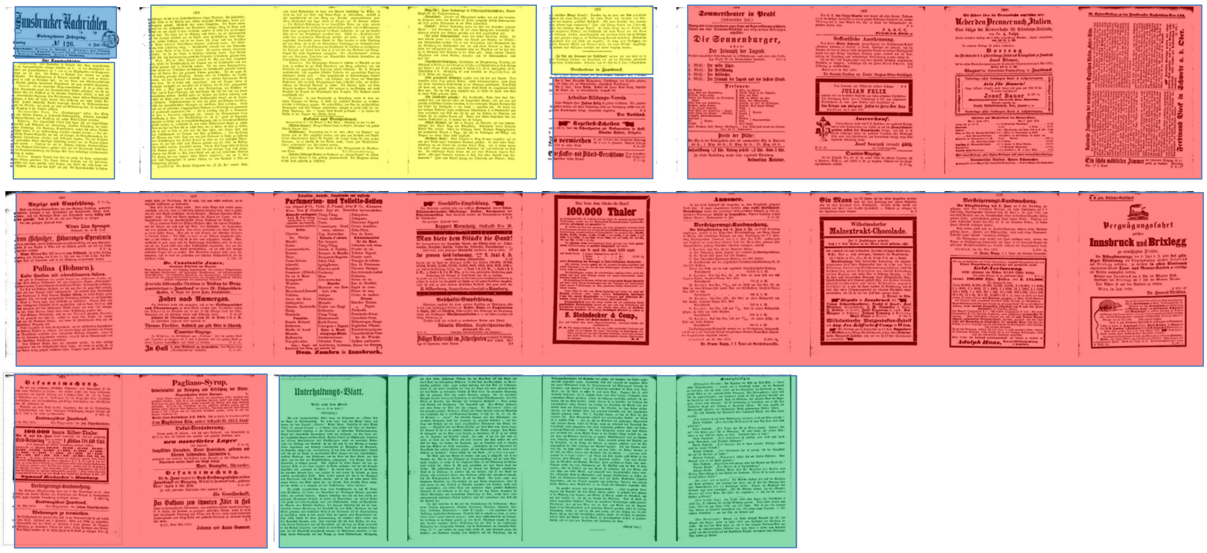


Figure 8: Metadata (blue), Information (yellow), Advertisements (red) and Entertainment (green)

5.7. *Text Types and Genres*

In a similar way as the five main classes of newspaper content, UIBK specifies text types or genres of news which appear in nearly every newspaper. Some typical examples are the “Editorial”, a (subjective) statement of the publisher, the editor in chief with regard to a recent event or the political, economic or cultural situation in general. Another example are book reviews, which were first introduced in the early 19th century and are until today one of the classical articles in the “Feuilleton” section of a newspaper. Other examples are Death notices, or Job and Real Estate offers.

It would be a highly interesting research field to explore the similarities of these genres with the techniques of computational linguistics. In combination with the layout of a newspaper and its divisions into sections, it can be expected that good results could be achieved with automated genre attribution.

In this case, UIBK will list some of the most prominent text types, being fully aware of the fact that the list would need much more investigation and research to be regarded as a comprehensive overview of all text types used in newspapers. This table shall therefore serve more as an example of how it could look like than an actual classification.

Category	First level	Second level	Definition
Information			
	News		The default class for all Newspaper Content Item in the Information category. Synonyms are “news article”, “article” or “story”.
		Lead Story	The main story, usually on the title page or on the title page of a section.
		Breaking News	The latest news.
		Background News	News that offer a wider view on a certain story, event or person.
		Reportage	Similar to background news, but written from a more personal or subjective perspective.
	Verbatim reports		
		Interviews	A news article consisting mainly of the verbatim report from an interview (with a well-known person or expert).
		Discussions	A news article consisting mainly of a verbatim report of a discussion of several persons.

		Speeches	A verbatim report from a speech (e.g. laudation or obituary).
	Biographical news		
		Portraits	A news article about a person.
		Anniversaries	An article on the occasion of an anniversary.
		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”.
	Factual news		The default category for all news that do not come as a narrative, but as lists or tables.
		Weather reports	Reports about the weather and weather forecast.
		Program news	Theatre, music hall, TV, radio programs. Most often organised as a list or table.
		Stock exchange rates	Stock exchange rates
		Railway tables	Timetables of railways, etc.
Opinion			
	Columns		A regular section of a newspaper or magazine devoted to a particular subject or written by a particular person.

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.
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
	Book review	A review about a (new) book.
	Theatre review	A review about a theatre performance.
	Concert review	A review about a concert.
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.
	Commentaries		Similar to editorials and columns commentaries are most often written by external authors, mostly experts.
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.
	Literary works		
		Serial novels	Serial novels are one of the most important classes and can be found in many historical newspapers from the 19th century onwards.
		Poems	A literary wok usually arranged in verses and strophes.
		Theatre plays	Sometimes parts of a theatre play are printed within newspapers.
		Essays	A piece of text mostly with philosophical considerations.

		Aphorism	A short piece of text expressing a sometimes surprising view on a given issue.
	Graphical works		
		Photos	
		Cartoons	
	Jokes		A piece of text to cause amusement or laughter, especially a story with a funny punch line.
	Games		
		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.
		Riddles	A question or statement intentionally phrased so as to require ingenuity in ascertaining its answer or meaning:
Advertisement			
	Advertisements		A piece of text or a figure used to encourage, persuade, or manipulate readers or to take or continue to take some action.
	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".
Family notices	Short notices related to families, such as marriage, birth of child, etc.
Death notices	Short and mostly formal notice on the death of a person.
Job offers	Announcement of open jobs.
Real estate offers	Announcements of available real estate objects.

6. Classification of Content Items

6.1. General considerations

Content Sections and Content Units are composed by pieces of text, but also photographs, tables and figures which can be characterised by their functional value together with their typical layout. These Content Items are described here as a comprehensive list.

In order to make the list as straightforward as possible, all Content Items were left out which are recorded anyway by libraries and which will play only a minor role in the deep structuring of newspapers. E.g. the newspaper title itself, the responsibility statement or the edition statement which can be found within the title section, need not to be mentioned here explicitly since they do – at least from this point of view – not play any substantial role in the structural mark-up of a newspaper.

6.2. List of Content Items

6.2.1. 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

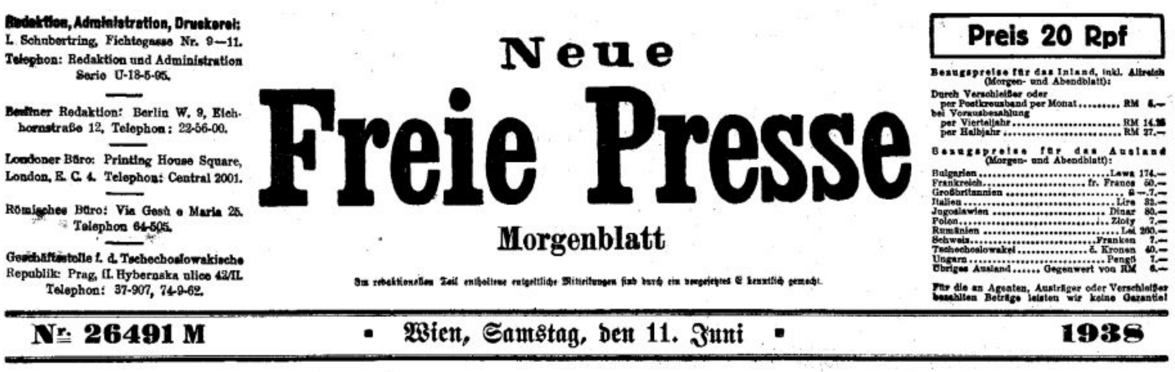
The title section can be regarded as a “shrunk” title page of a book. Usually the descriptive metadata contained in the title section, such as name of the newspaper, edition, publisher, issue date, etc. are captured as part of a library catalogue or in beforehand of the scanning process. As already indicated above from the point of view of structural metadata the title section is just noise for information retrieval and should therefore to be excluded from full-text search or text mining operations.

Automated Capturing

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

Examples

A classical title section from the “Neue Freie Presse”, 11th June 1938 and from the French newspaper “L’ Intransigeant” from 1884.



Redaktion, Administration, Druckerei:
I. Schubertring, Fichtegasse Nr. 9—11.
Telephon: Redaktion und Administration
Serie U-18-5-95.

Berliner Redaktion: Berlin W. 9, Eichhornstraße 12, Telephon: 22-56-00.

Londoner Büro: Printing House Square, London, E. C. 4. Telephon: Central 2001.

Römisches Büro: Via Gesù e Maria 25.
Telephon 64-505.

Geschäftsstelle f. d. Tschechoslowakische Republik: Prag, II. Hybernska ulice 42/II.
Telephon: 37-907, 74-9-62.

Neue
Freie Presse
Morgenblatt

Im redaktionellen Teil enthaltene entgeltliche Mitteilungen ſind durch ein vorgeſetztes E kenntlich gemacht.

Preis 20 Rpf

Bezugspreise für das Inland, inkl. Altreich
(Morgen- und Abendblatt):

Durch Verschleißer oder
per Postkreuzband per Monat .... RM 5.—
bei Vorausbezahlung
per Vierteljahr .... RM 14.25
per Halbjahr .... RM 27.—

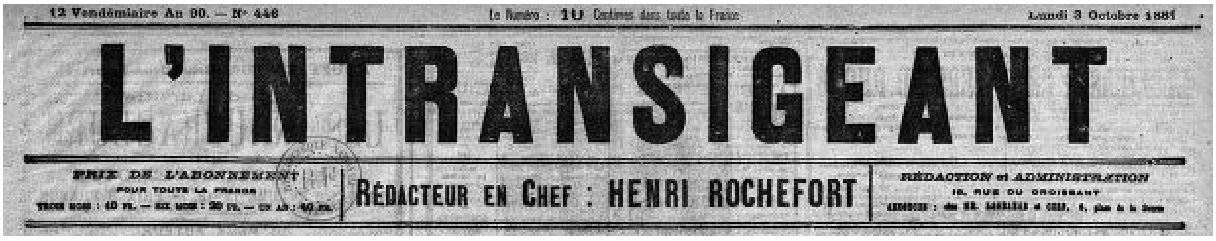
Bezugspreise für das Ausland
(Morgen- und Abendblatt):

Bulgarien	Lewa	174.—
Frankreich	fr. Francs	50.—
Großbritannien	£	—.7.—
Italien	Lire	32.—
Jugoslawien	Dinar	80.—
Polen	Zloty	7.—
Rumänien	Lei	200.—
Schweiz	Franken	7.—
Tschechoslowakei	č. Kronen	40.—
Ungarn	Pengö	7.—
Übriges Ausland	Gegenwert von RM	6.—

Für die an Agenten, Austräger oder Verschleißer bezahlten Beträge leisten wir keine Garantie!

Nr. 26491 M ▪ Wien, Samstag, den 11. Juni ▪ 1938

Figure 9: Title section, 1938



12 Vendémiaire An 90. — Nº 446 Le Numéro : 10 Centimes dans toute la France Lundi 3 Octobre 1881

L’INTRANSIGEANT

PRIX DE L’ABONNEMENT
POUR TOUTE LA FRANCE
TROIS MOIS : 10 FR. — SIX MOIS : 20 FR. — UN AN : 40 FR.

RÉDACTEUR EN CHEF : HENRI ROCHEFORT

RÉDACTION et ADMINISTRATION
18, RUE DU CROISSANT
ANNONCES : chez MM. LAGRANGE et CERF, 6, place de la Bourse

Figure 10: Title section, 1884

6.2.2. Running title

Synonyms

Header, column title.

Definition

The running title 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

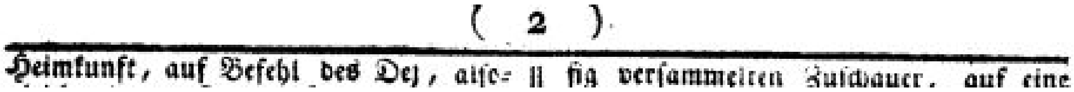
The functional value of the running title is to provide the user a quick orientation which newspaper issue he is actually reading (number, date) and maybe some information on the content of the page (e.g. if a section heading is included). The benefit of running titles for deep structuring is rather 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 detail the content of a page.

Automated Capturing

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

Examples

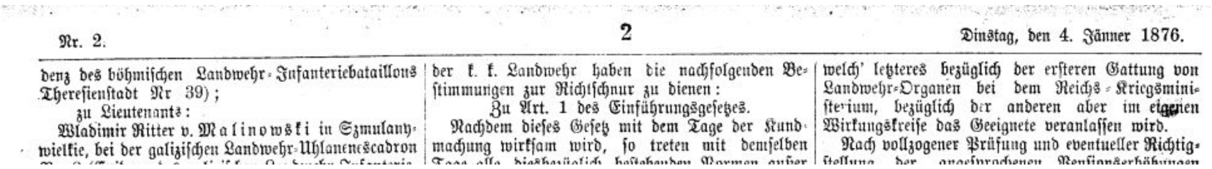
The “Wiener Zeitung” introduced its first running title with the first issue of the year 1784. It contained for more than 100 years just the page number, from 1876 onwards, also the issue number, title, and date.



(2).

Heimkunft, auf Befehl des Dej, alſo= ll ſig verſammelten Zuſchauer. auf eine

Figure 11: Running title with the page number only, 1784



Nr. 2. 2 Dinstag, den 4. Jänner 1876.

denz des böhmiſchen Landwehr-Infanteriebataillons Thereſienſtadt Nr 39);

zu Lieutenants:

Wladimir Ritter v. Malinowski in Szmulany-wielkie, bei der galiziſchen Landwehr-Uhlanenescadron

der k. k. Landwehr haben die nachfolgenden Beſtimmungen zur Richtſchnur zu dienen:

Zu Art. 1 des Einführungsgeſetzes.

Nachdem dieſes Geſetz mit dem Tage der Kundmachung wirkſam wird, ſo treten mit demſelben

welch' letzteres bezüglich der erſteren Gattung von Landwehr-Organen bei dem Reichs-Kriegsminiſterium, bezüglich der anderen aber im eigenen Wirkungskreiſe das Geeignete veranlaſſen wird.

Nach vollzogener Prüfung und eventueller Richtig-

Figure 12: Classical running title with issue number, page number and date, 1876

6.2.3. Heading

Synonyms

Head, main title, title.

Definition

Articles (Content Units) within a newspaper usually come with a title indicating briefly the content of the article. Large articles may have several titles, such as a top heading, a heading, a sub-heading and several inside-headings.

Value

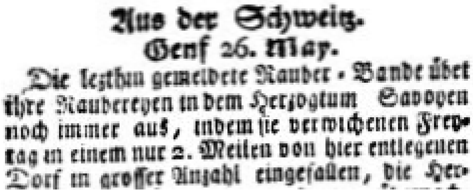
The functional value of headings is twofold: First, to mark the beginning of a Content Unit by providing in some way an “eye catcher” which is in the easiest case a word in bold, or in brackets, or in later times a word with a large font size. Second, to attract the attention of a reader to actually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

Automated capturing

In modern newspapers the capturing of headings can be done with good success since the different font size compared to the running text is a very good indicator. 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 there are several ways to detect the headline in an indirect way, by utilizing the background knowledge which is set out here. 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.

Examples

It took until the middle of the 19th century that newspapers started to indicate the start of a news item with a specific heading. Before that time it was the place of origin (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.



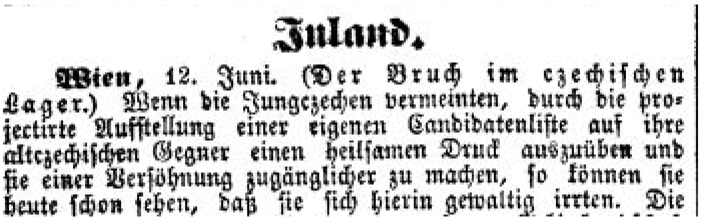
Aus der Schweiz.

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 13: Section heading and Coverage Note spatial

100 years later, there is still a high similarity: 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 also in spaced letters and put into brackets in order to separate it from the actual running text.

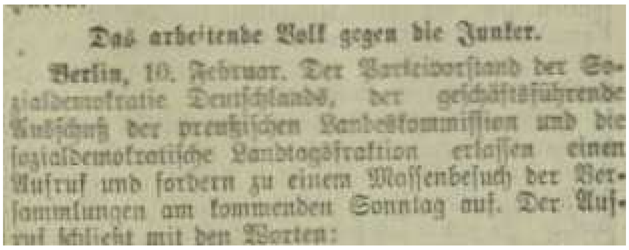


Inland.

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 14: Die Presse, 1874

It takes until the 20th century that the modern form is found with a heading indicating the start of the news and a coverage note with additional information. Heading: “Das arbeitende Volk gegen die Junker”, “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 15: Classical start of news, Arbeiterwille, 1910

6.2.4. Sub-heading

Definition

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

Functional Value

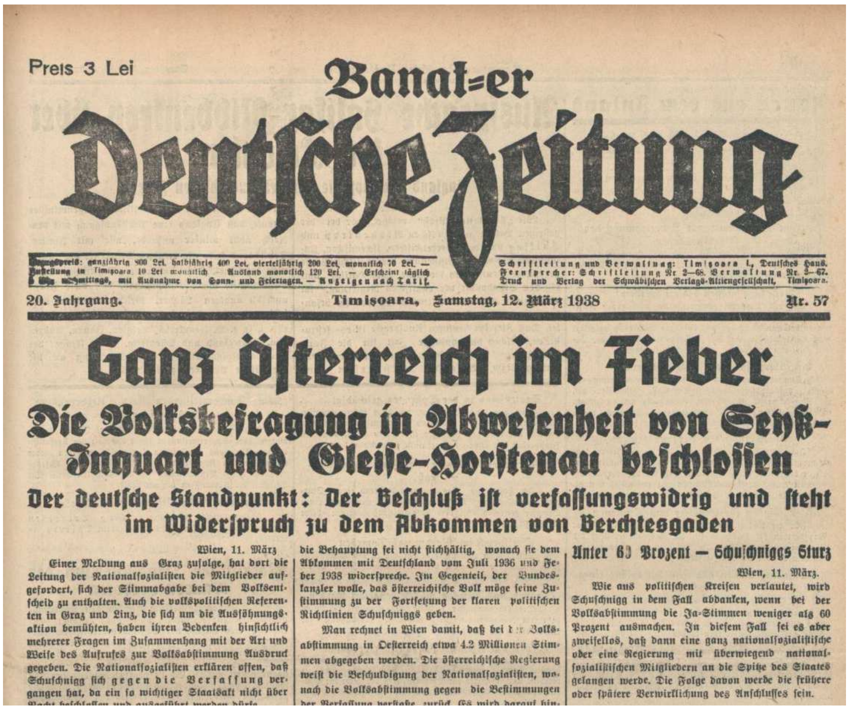
The value is high since the content of the news is often explained in more detail.

Automated Capturing

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

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:



Preis 3 Lei

Banat=er
Deutſche Zeitung

Bezugspreis: ganzjährig 800 Lei, halbjährig 400 Lei, vierteljährig 200 Lei, monatlich 70 Lei. — Zuſtellung in Timișoara 10 Lei monatlich — Ausland monatlich 120 Lei. — Erſcheint täglich 5 Uhr nachmittags, mit Ausnahme von Sonn- und Feiertagen. — Anzeigen nach Tarif.

Schriftleitung und Verwaltung: Timișoara I, Deutſches Haus. Fernſprecher: Schriftleitung Nr 2—68. Verwaltung Nr. 2—67. Druck und Verlag der Schwäbiſchen Verlags-Aktiengeſellſchaft, Timișoara.

20. Jahrgang. Timișoara, 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 Nacht beſchloſſen und ausgeführt werden dürfe. die Behauptung ſei nicht ſtichhältig, wonach ſie dem Abkommen mit Deutſchland vom Juli 1936 und Feber 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 der Verfaſſung verſtoße, zurück. Es wird darauf hin-

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

6.2.5. Inside-heading

Synonyms

Sub-heading.

Definition

Larger articles are sometimes structured with headings directly within the text body. In contrast to highlighted words or phrases which do not disturb 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 letters-space, 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 such as bold, or letter spaced), if this piece interrupts the flow of text then it is an inside-heading.

Functional Value

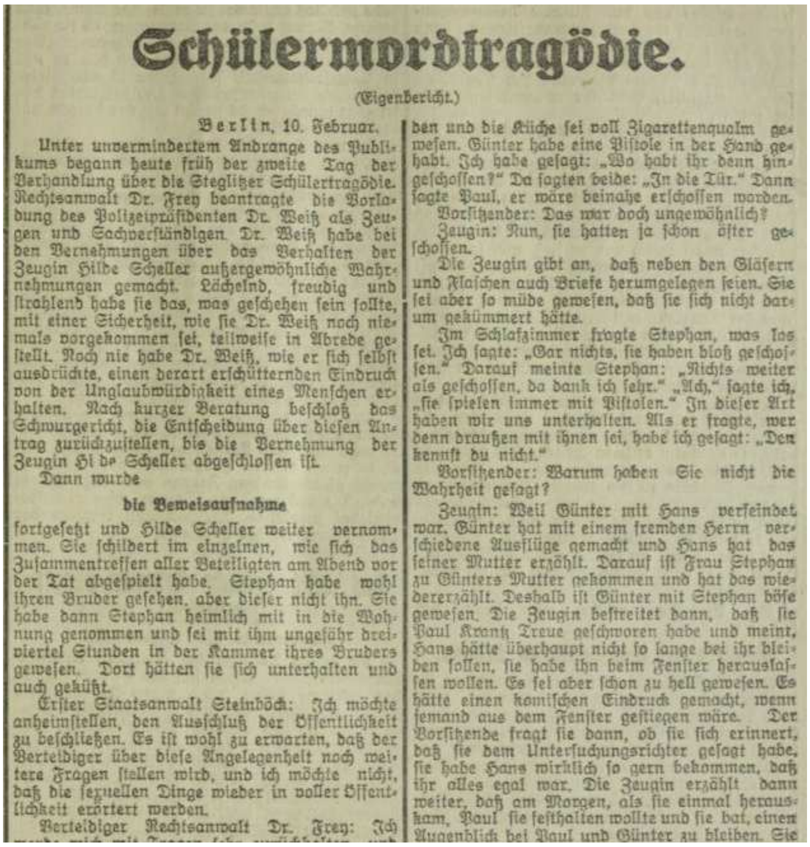
Since inside-headings are embedded particles within the text they “destroy” the running text. This may have some negative affect if the text is parsed, e.g. with natural language parsers which is a usual step in the workflow for information extraction.

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.

Example

In the following, a good example of a pure highlighting which may look like an inside-heading is shown.



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 Hi de 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 17: Arbeiterwille, 11th February 1928

6.2.6. 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.

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.

6.2.7. Lead

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 Content Items. 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.

6.2.8. 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.

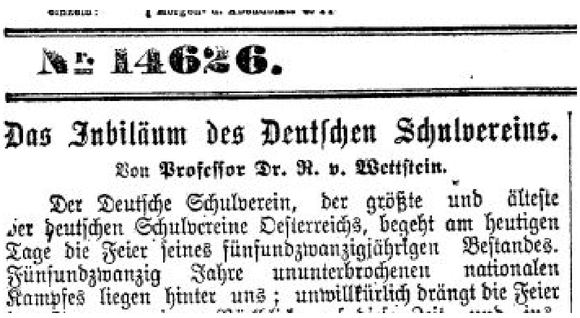
Again for automated processing, copyright notes can be regarded as embedded particles and are somehow disturbing the running text which may be subject to automated processing.

Automated Capturing

Tthere are rather strict rules for each newspaper where these copyright notes appear and also the text itself is rather strict and follows simple rules. In the case of individual authors once can expect that a “Von” or “By” is used as suffix of this note. The chance to detect this content item automatically is therefore high.

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.



Nr. 14626.

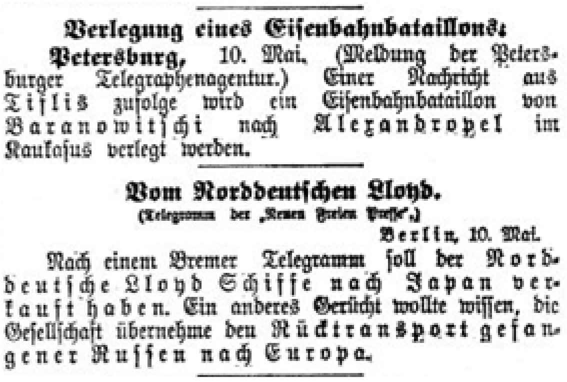
Das Jubiläum des Deutſchen Schulvereins.

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 18: Copyright note, 1905

Here one can find three 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 19: 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”.)

6.2.9. Coverage note spatial

Synonyms

Place name.

Definition

Especially news articles commonly indicate the location of a story right at the beginning of the item. Coverage notes are defining the location (or time) 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 Content Units.

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 deep structuring is very 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.

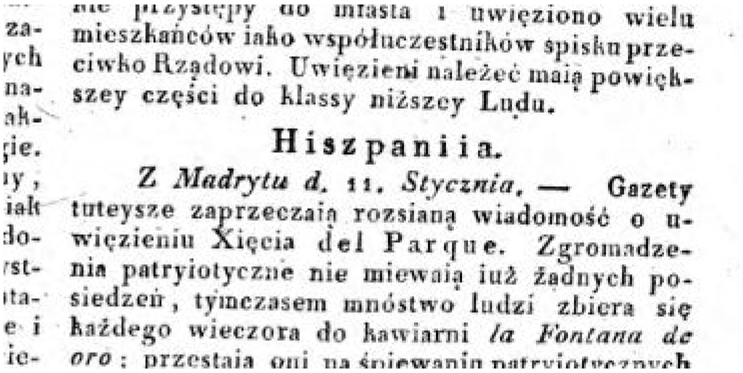
From the historical point of view Coverage notes spatial are very similar to section headings which are consisting just of the place name of a specific section.

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.

Examples

Gazeta Lwowska (Lemberger Zeitung / Poland), 16. February 1821. The complete coverage note includes the place and detailed date of the event: Madrytu, and the 11th of January (!).



nie przystępy do miasta i uwięziono wielu mieszkańców iako współuczestników spisku przeciwko Rządowi. Uwięzieni należeć maią powiększey części do klassy niższey Ludu.

Hiszpaniia.

Z Madrytu d. 11. *Stycznia.* — Gazety tuteysze zaprzeczaią rozsianą wiadomość o uwięzieniu Xięcia del Parque. Zgromadzenia patryiotyczne nie miewaią iuż żadnych posiedzeń, tymczasem mnóstwo ludzi zbiera się każdego wieczora do kawiarni *la Fontana de oro*; przestaią oni na śpiewaniu patryiotycznych

Figure 20: Example of a coverage note, 1821

6.2.10. 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.

Functional Value

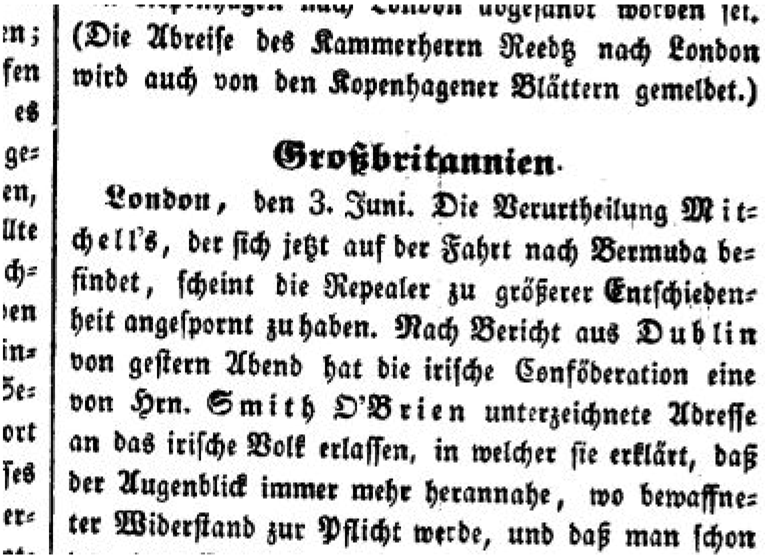
With the coverage note temporal the reader gets detailed information about the “age” of a news. This was especially important at former times when the transportation of news took several days or – if they came from abroad – even longer. In contrast to the coverage note spatial, the information value 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 of higher value.

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 since their repertoire is very limited and they 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.

Examples

The example 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.



Kopenhagen nach London abgeſandt worden ſei. (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 21: An example of a coverage note, 1848

6.2.11. Paragraph

Definition

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

Functional Value

The functional value of paragraphs is to structure a longer piece of text into suitable parts. Usually a paragraph covers one thought or one distinct part of the narrative or message. The value is very high, since a paragraph can be seen as a basic building block of any text. The correct separation of the text into paragraphs is important for many reasons, especially for any kind of Natural Language Processing where tree- and part-of-speech taggers are used. They will in many cases rely on the correct start and end of sentences.

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.

6.2.12. 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 a significant increase of illustrations within newspapers can be seen, 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.

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

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.

6.2.14. 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 are taken as named entities and linked to corresponding resources, such as library catalogues.

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.

6.2.15. Continuation note

Definition

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

Functional 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 though, since it will connect dislocated pieces of a single 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.

6.2.16. 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. UIBK includes 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 “emphasiser” 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 in another place.

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.

6.2.17. 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

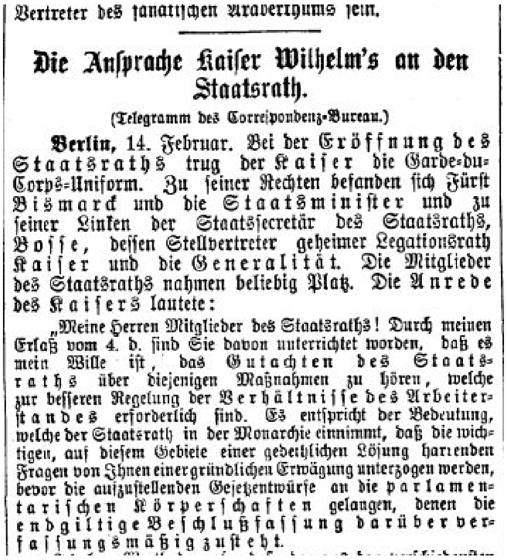
In order to emphasise the importance of the message of a person or an institution a piece of text was included as a verbatim quote. They are usually marked with their layout.

Automated Capturing

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

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. 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 22: Verbatim note, indicated by smaller font size, 1890

6.3. Representation in MODS

Content Items are on the one hand used to support the user in navigating through a newspaper issue, but they are also containing some “meta-information” which can be regarded as “descriptive metadata”. If the Content Items are separated into two main groups, those which directly belong to the running text of a news and those which provide additional information but are not part of the running text, one gets a meaningful compilation.

Content Items forming the main part of the running text	Content items which provide additional information
Paragraphs	Title section
Illustrations, tables, charts	Running title
Captions	Headings (including sub-heading, top-heading, inside heading)
Verbatim	Copyright note
Lead	Coverage notes (spatial and temporal)

	Continuation note

Looking at those Content Items which provide additional “meta-information”, a simple match can be made between the MODS schema and the list of Content Items. In this way, a detailed representation of all newspaper Content Units and Content Sections within the MODS schema is obtained.

Content Item	Type of Metadata	Representation in MODS
Title section	The title section contains a lot of metadata which are recorded in a library catalogue, such as issue number, name of the newspaper, publication place, etc. But as discussed above, it will not be necessary to represent and capture Title sections in detail, since there is information anyway in beforehand.	Not applicable.
Running title	The same is true with running titles, with one exception: If the running title also includes a “subject heading” indicating the content of the page. E.g. “Sports”.	MODS Subject.

Headings of Content Sections	Since headings of content sections are repeated across issues they can be seen as “subject” headings, indicating that the Content Units within the section fall under a specific category. E.g. all news under the heading “News from abroad”. In many cases MODS Subject – Topic could be used, also “Geographical” and “Temporal”.	MODS Subject
Headings of Content Units	Headings of Content Units are unique and individual and go therefore under the “Title Info” of the MODS schema.	MODS Title
Copyright note	The copyright note indicates the source of information and can therefore go under “Name”. MODS allows here also to record if it is an individual creator or a corporate one, which will be useful for cases where just the newspaper is mentioned as copyright holder.	MODS Name
Coverage notes (spatial and temporal)	Coverage notes indicate the subject (similar to content sections) and can be recorded as MODS Subject,	MODS Subject

	Geographical and Temporal.	
Continuation note	The continuation note may be regarded as a distinct part of a news, than the two items could be linked together with the MODS Part element. Or – which will be in many cases more appropriate – the continuation note will not be recorded at all since it does in many cases not provide any additional information.	MODS Part

In addition to this table it is obvious that for recording text types, the MODS Genre field could be used, e.g. “Book review”.

7. STRUCTIFY

7.1. General

STRUCTIFY is a JAVA tool utilising the Standard Widget Toolkit (SWT) for displaying, rendering and configuring METS files, respectively ENMAP files and their associated image and text files. It was developed by UIBK. It is available for free for download from the FTP server of UIBK. A detailed description can be found here:

http://dbis-halvar.uibk.ac.at/dokuwiki/doku.php?id=main:structify

The main idea of STRUCTIFY is to display images and the full-text of digitized newspapers (and other formats).

One of the main features of STRUCTIFY is its flexibility which is based on a sophisticated “handler” system. In this way it can be adapted to all kinds of METS and ALTO formats with a minimum of effort. For example, there is a handler to open the ENMAP simple and one for the ENMAP extended version. Moreover, a handler which supports the CCS output format was developed in year 3 to give the libraries the opportunity to view the results from the OLR process of CCS. Since some libraries may also be interested to use the ABBYY FineReader XML format for storing textual data, there is also a handler to work with these XML files. In addition it is also possible to work without any initial METS at all – and just import images plus OCR files. This way someone could use Structify to create ENMAP (simple or extended) with the help of the tool and use the output as a specification for e.g. a service provider or internal discussions.

Apart from the handler system, also several widgets have been introduced to shorten some workflow steps. Again the wizard system is designed in a generic way, so that specific tasks can be supported easily.

7.2. ENMAP Viewer

A first and very simple purpose of STRUCTIFY is to use it as a simple viewer program for the files produced in the Europeana Newspaper project. It is able to directly load the ENMAP METS together with the image and ALTO files and to display regions and text. It has also been adapted to work with the METS and ALTO profile from CCS GmbH (Content Conversion Specialist) which is slightly different from the ENMAP enhanced format.[9]

Several libraries, among them the British Library and the Bibliothèque nationale de France, are now using STRUCTIFY for (visual) quality control of METS files.

7.3. Ground Truth and Quality Assessment

One of the key factors for successful digitisation and enrichment projects is to translate the demands of libraries and humanities scholars into technical requirements which need to be fulfilled by service providers. Usually such requirements come in a written form and may be rather vague. Since the final end-product delivered by the service providers are highly complex and – for human beings – rather hard to understand XML files, a direct evaluation of the results is often only possible with the tools provided by the service providers themselves.

This “translation process” can be significantly simplified if the requirements are already exposed in the final format and are visually accessible to those people who have the domain knowledge about the content of historical newspapers. All the detailed decisions which need to be made if a library wants to follow above suggestions for a “deep structuring” can be laid out directly with STRUCTIFY by non-technical people. In this way the tool may play an important role as a “communication” tool between libraries/humanities scholars and technology providers.

Due to the fact that STRUCTIFY provides the exact ENMAP output, also technical people are able to directly understand how the encoding shall take place.

Strongly connected with the generation of Ground Truth STRUCTIFY can also be used as a quality assessment tool. Assuming that – based on Ground Truth produced by a library – a service provider will deliver large amounts of structured newspapers, a defined quality assessment process may take place. In contrast to the OCR assessment tools from USAL which were developed in Work Package 3 there is currently no automated process included in STRUCTIFY which would allow a direct evaluation of the delivered product against reference data. On the other hand, based on some random samples, a library may organise such a quality assessment process in a rather simple way, by just viewing some files and

[9] The main reason for this inconsistency is that in the work plan it was neither foreseen nor possible to adapt the CCS METS format according to the ENMAP enhanced profile.

recording errors either with a separate tag (as it can be defined in STRUCTIFY) or in an external list.

7.4. *Training Data*

As already mentioned deep structuring of newspapers will only be possible with the strong support of technologies from Pattern Recognition, Natural Language Processing and similar research fields. Most of these algorithms are based on machine learning techniques which require so-called "training data". As a rule of thumb it can be stated that the more training data are available, the better the results will be. In fact, the progress achieved in similar pattern recognition tasks, such as Speech Recognition, Computer Vision or Online Handwritten Text Recognition are mainly based on the improved availability and quantity of said training data. The drawback of this approach is that the generation of training data is cumbersome and requires a lot of manual labour.

When taking these considerations into account, a project plan for the deep structuring of a large amount of newspapers may look like this:

1. Set up requirements
   Based on STRUCTIFY and ENMAP, several examples are produced for all structural features (Content Sections, Units and Items, Text Types, etc.) which shall be detected in an automated process.

2. Use expert system for generating basic data
   By a rule-based approach, expert and domain knowledge can be heavily utilised to automatically detect all structural features as they were defined in the first step. It can be expected that in some cases this rule based approach will lead to rather good results (e.g. the detection of title sections, in other cases it may be very erroneous).

3. Training Data Generation
   The actual generation of training data may take place in this third round and is based on the results produced by the (rule based) expert systems. Only the errors of the expert system need to be corrected - which can also be done with STRUCTIFY. In this way a large number of data will be available for the next step.

4. Machine Learning Approaches
   Based on a significant number of data, machine learning methods may now be applied. It can be expected that if indeed all the structural features which are mentioned above shall be detected in an automated way, some ten-thousands of reference pages will be necessary and therefore several iterations may be applied. These iterations may be continued as long as there are significant improvements in the accuracy.

7.5. Digital Humanities Tool

Independently of the fact how good the actual results of the automated processing for deep structuring will be – they never will be perfect. Therefore the need for some final correction process will always remain. Given the large number of newspaper pages it is illusionary to believe that the complete process of deep structuring may be carried out or be financed by libraries. The involvement of user groups who are interested in receiving improved results is by all means necessary on the long term. It is therefore expected that not only the correction of OCR text will be done by volunteers and humanities scholars, but also the correction of structural features. Whereas the correction may be done on the basis of web-interfaces for simple features, such as applying text types to Content Units, for operations which will require a more complex rendering of the data, again STRUCTIFY may play a significant role.

7.6. STRUCTIFY Screenshot

In order to provide a short impression of the tool, a screenshot of the tool is provided here. The tool consists of five main areas.

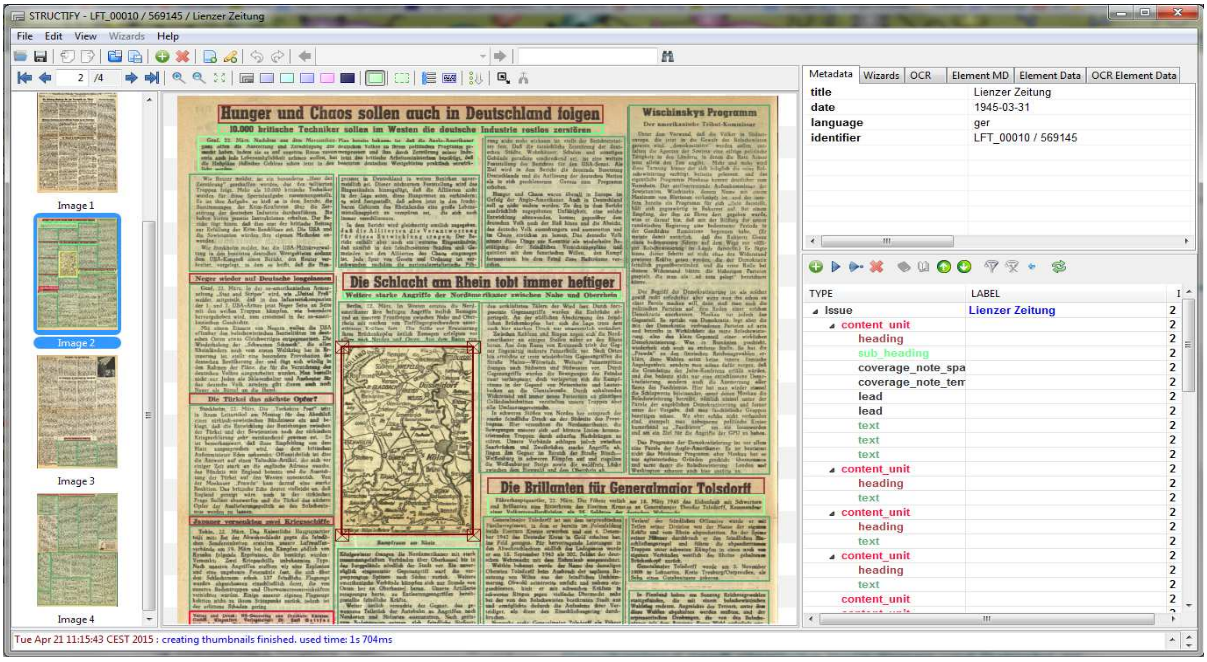
1. The menu bar on top

2. A thumbnail view of the document on the very left hand side.

3. An image canvas in the centre of the screen where the actual page and the raw segmentation (coming from the OCR engine) as well as the logical structure of the document is displayed.

4. A tree map at the bottom right hand side displaying the Structural Map of the METS file and therefore showing the actual structuring.

5. A metadata area on the top right hand side where on the one hand parts of the descriptive metadata section of the METS file are displayed as well as specific wizards can be utilised to speed up the tagging process.


File
Edit
View
Wizards
Help
2
/4
Image 1
Image 2
Image 3
Image 4
Hunger und Chaos sollen auch in Deutschland folgen
10.000 britische Techniker sollen im Westen die deutsche Industrie restlos zerstören
Wischinskys Programm
Der amerikanische Tribut-Kommissar
Neger wieder auf Deutsche losgelassen
Die Türkei das nächste Opfer?
Die Schlacht am Rhein tobt immer heftiger
Weitere starke Angriffe der Nordamerikaner zwischen Nahe und Oberrhein
Die Brillanten für Generalmajor Tolsdorff
Japaner versenkten zwei Kriegsschiffe
Metadata
Wizards
OCR
Element MD
Element Data
OCR Element Data

STRUCTIFY - LFT_00010 / 569145 / Lienzer Zeitung

title	Lienzer Zeitung
date	1945-03-31
language	ger
identifier	LFT_00010 / 569145

TYPE	LABEL	I
Issue	Lienzer Zeitung	2
content_unit		2
heading		2
sub_heading		2
coverage_note_spa		2
coverage_note_ten		2
lead		2
lead		2
text		2
text		2
text		2
content_unit		2
heading		2
text		2
content_unit		2
heading		2
text		2
content_unit		2
heading		2
text		2
content_unit		2

Tue Apr 21 11:15:43 CEST 2015 : creating thumbnails finished. used time: 1s 704ms

Figure 23: STRUCTIFY Screenshot

8. Summary

A short summary of the most important concepts, rules and mappings which are provided in the following section:

Deep Structuring

- Full informational capture of structural features of a newspaper issue
- Granular approach

Content Units

- The classical entity of newspapers, such as articles, advertisements, classified advertisements
- Clearly distinguishable from neighbouring pieces by its content
- May include other Content Units if it is a complex or large piece of text

Content Sections

- Sections provide a rough structure of a newspaper issue
- Sections serve as placeholders for content units falling under a given category
- Sections are always repeated in various issues and can therefore only be identified across the borders of a single newspaper issue

Content Items

- Content Items are the building blocks of Content Units and Content Sections
- They are mainly defined by their functional value and their layout
- Content Items are important for the internal structure of the content.
- Content Items also contain some very specific descriptive metadata which can be exploited for metadata recording

Classification of Content Items into classes

- Main classes are: Title section, running title, headings (top-, sub-, inside-heading), copyright note, coverage note (spatial and temporal), continuation note, paragraph, illustration, table, list, caption, lead, verbatim note and summaries

Classification of content into five main classes

- Information (news)
    - The classical news content with text (and illustrations) about recent events built around the five “Ws”: Who, What, Where, When and Why
- Advertisement, including classified advertisement
    - External content ranging from official announcements to commercial advertisements
- Entertainment
    - All kinds of arts and literature. Prominent examples are the series novels in the 19th and cartoons in the 20th century
- Opinion

- A personal reflection or standpoint. Started with book reviews in the early 19th century and continued with “editorials” and “commentaries” as main examples
- Metadata
  - Some content pieces which provide information about the newspaper itself

Classification of Content Units into text types

- Extended list of genres

Matching Content Items with MODS

- Running title → Not recorded, except section heading → MODS Subject
- Headings of Sections → MODS Subject
- Headings of Units → MODS TitleInfo
- Copyright note → MODS Name
- Coverage notes (spatial and temporal) → MODS Subject (geographical, temporal)
- Continuation note → MODS Part

ENMAP format

- ENMAP simple: image and OCR files on page basis
  → used to deliver information packages to the Europeana Newspapers application: http://www.theeuropeanlibrary.org/tel4/newspapers
- ENMAP enhanced → experimental format for describing mainly the structural features of historical newspapers

STRUCTIFY

- A free tool for displaying, rendering and generating ENMAP enhanced files on the basis of ENMAP simple

9. ENMAP – Profile

9.1. Examples

Part of this deliverable are some example issues and pages for ENMAP based on a manual tagging of content pieces. These examples are also available for download from the Europeana Newspapers website and can be displayed with STRUCTIFY.

9.2. ENMAP Profile

This section provides a detailed XML profile description of ENMAP (Europeana Newspaper METS ALTO Profile). It describes how to use the elements and attributes from the Metadata Encoding and Transmission Standard for the purpose of digitised newspapers.

Note: METS Elements and Attributes not covered by this document are currently not used and therefore not mentioned in this profile description.

The XML Prolog

Defines the used XML Version and the used character encoding, the preferred XML Version is 1.0 and the preferred encoding is UTF-8.

The METS-Root Element: mets

Namespace: http://www.loc.gov/METS/

Description: This is the main container and contains all other metadata sections (METS header, descriptive metadata, administrative metadata, file section, structure map) and all the namespace definitions for all the used metadata standards.

Repeatable: no

content/childs: metsHdr, dmdSec, amdSec, fileSec, structMap

Attributes:

xmlns:mets	Defines the namespace of the METS container http://www.loc.gov/METS/	REQUIRED

xmlns:xsi	The XML-Schema-Instance namespace definition, needed for XML validation http://www.w3.org/2001/XMLSchema-instance	REQUIRED
xmlns:mix	The National Information Standards Organization namespace. NISO mix is used to store administrative metadata for the contained files http://www.loc.gov/mix/v20	REQUIRED
xmlns:mods	Namespace of the used descriptive metadata standard MODS http://www.loc.gov/mods/v3	REQUIRED
xmlns:xlink	XLink Namespace, used and referenced by the METS Schema http://www.w3.org/1999/xlink	REQUIRED
PROFILE	Used to determine the XML profile as an European Newspaper METS/ALTO Profile, always set to: ENMAP	REQUIRED
OBJID	Contains an unique identifier for the dataset, value can be any string	OPTIONAL

The METS Header Element: metsHdr

Namespace: http://www.loc.gov/METS/

Description: The METS Header contains the records status and the modification dates, as well as a list of agent and their role assigned to this Mets document.

Repeatable: no

content/childs: agent

Attributes:

RECORDSTATUS	Contains a String representing the actual state, e.g.: SUBMITTED	REQUIRED
CREATEDATE	Contains the date on which the METS document was created. Used format is a XMLDateTime 2014-02-18T12:28:21	REQUIRED
LASTMODDATE	Holds the last date on which the document was modified. Used format is a XMLDateTime 2014-02-18T12:28:21	REQUIRED

The METS agent-Element: agent

Namespace: http://www.loc.gov/METS/

Description: defines an agent and its role on this newspaper issue

Repeatable: yes

Content/childs: name

Attributes:

ROLE	Defines the role of the agent, valid values are adopted from the METS Schema e.g. CREATOR, CUSTODIAN, ….	REQUIRED
TYPE	Determines the agent type ORGANIZATION, INDIVIDUAL or OTHER	REQUIRED

The METS agent name-element: name

Namespace: http://www.loc.gov/METS/

Description: contains the agent name

Repeatable: no

Content/childs: a string (TextNode) to identify the agent

The METS descriptive metadata section-element: dmdSec

Namespace: http://www.loc.gov/METS/

Description: each descriptive metadata section contains exactly one metadata set which can be referenced from the structure map. The ENMAP Profile awaits at least one record set referenced from the structMap root div, which should contain metadata about the whole newspaper or newspaper issue.

Repeatable: yes

Content/childs: mdWrap

Attributes:

ID	XML ID used by the structMap to reference this metadata set.	REQUIRED

The METS administrative metadata section element: amdSec

Namespace: http://www.loc.gov/METS/

Description: the administrative metadata section contains a list of metadata sets which are referenced from the file section to hold additional image metadata like width, height, compression schema and others.

Repeatable: no

Content/childs: techMD

Attributes:

ID	XML ID used to define this section as a technical metadata section TECHMD	REQUIRED

The METS technical metadata-element: techMD

Namespace: http://www.loc.gov/METS/

Description: this element is referenced from the file section and holds image metadata

Repeatable: yes

Content/childs: mdWrap

Attributes:

ID	XML ID used from the file section to reference this metadata set	REQUIRED

The METS metadata wrapper-element: mdWrap

Namespace: http://www.loc.gov/METS/

Description: used to specify the type of the wrapped metadata set.

Repeatable: no

Content/childs: xmlData

Attributes:

MDTYPE	defines the metadata type, in case of descriptive metadata MODS is used, and in case of administrative metadata NISOIMG is used	REQUIRED

The METS xml data wrapper-element: xmlData

Namespace: http://www.loc.gov/METS/

Description: used to specify the type of the wrapped metadata set as a XML fragment.

Repeatable: no

Content/childs: mods, mix

MODS

The MODS Root-Element: mods

Namespace: http://www.loc.gov/mods/v3

Description: root element of the MODS xml record.

Repeatable: no

Content/childs: all MODS root-childs are allowed here, for a complete list see http://www.loc.gov/standards/mods/

Below are some important MODS elements which are used in ENMAP profile in the simple as well as the extended version. These elements were valuable for later data processing and hence recommended as best practice.

The MODS titleInfo-Element: titleInfo

Namespace: http://www.loc.gov/mods/v3

Description: provides the title information

Repeatable: yes

Content/childs: title

titleInfo		REQUIRED

The MODS titleInfo title-Element: title

Namespace: http://www.loc.gov/mods/v3

Description: contains the newspaper title

Repeatable: yes

Content/childs: a string (TextNode) as title representation

title	Title of the newspaper	REQUIRED

The MODS originInfo-Element: originInfo

Namespace: http://www.loc.gov/mods/v3

Description: used to hold the origin information

Repeatable: no

Content/childs: dateIssued

The MODS originInfo dateIssued-Element: dateIssued

Namespace: http://www.loc.gov/mods/v3

Description: contains the publication date of the issue

Repeatable: no

Content/childs: a string (TextNode) to show the date

Attributes:

encoding	*w3cdtf* - This value is used for the profile of ISO 8601 that specifies the following date pattern: YYYY-MM-DD	REQUIRED
keyDate	yes - This value is used so that a particular date may be distinguished among several dates. Thus for example, when sorting MODS records by date, a date with	REQUIRED

	keyDate="yes" would be the date to sort on. It should occur only for one date at most in a given record.	

The MODS language-Element: language

Namespace: http://www.loc.gov/mods/v3

Description: used to hold the language and script type information

Repeatable: yes

Content/childs: languageTerm, scriptTerm

The MODS language languageTerm -Element: languageTerm

Namespace: http://www.loc.gov/mods/v3

Description: contains the language

Repeatable: yes

Content/childs: a string (TextNode) to identify the language

Attributes:

type	Type is either code or text	REQUIRED
authority	Enumeration of different language codes, eg. iso639-2b, rfc4646	REQUIRED

The MODS language scriptTerm -Element: scriptTerm

Namespace: http://www.loc.gov/mods/v3

Description: contains the script type

Repeatable: yes

Content/childs: a string (TextNode) to identify the script

Attributes:

type	Type is either code or text	REQUIRED
authority	Enumeration of different script codes, eg. iso15924	REQUIRED

The MODS identifier-Element: identifier

Namespace: http://www.loc.gov/mods/v3

Description: contains external identifiers

Repeatable: yes

Content/childs: a string (TextNode) to identify (external) representations of the same dataset

Attributes:

type	The identifier type	REQUIRED

The MODS accessCondition-Element: accessCondition

Namespace: http://www.loc.gov/mods/v3

Description: contains access information

Repeatable: yes

Content/childs: a string (TextNode) to define the access condition

Attributes:

type	The access type resp. access authority	REQUIRED

NISO

The NISO MIX Root-Element: mix

Namespace: http://www.loc.gov/mix/v20

Description: root element of the NISO MIX xml record.

Repeatable: no

Content/childs: all MIX root-childs are allowed here, for a complete list see http://www.loc.gov/standards/mix

The METS file section element: fileSec

Namespace: http://www.loc.gov/METS/

Description: The file section contains all files assigned to this newspaper, ordered into specified groups.

Repeatable: no

content/childs: fileGrp

The METS file group element: fileGrp

Namespace: http://www.loc.gov/METS/

Description: file groups are used to keep track on the different file types assigned to a digitized newspaper issue. Groups can again contain groups. Following file grouping is foreseen for ENMAP documents:

- ImageGroup
  - OCRMasterFiles: contains the original scans/images
  - ViewingFiles: this group can contain downscaled images used for fast displaying
- TextGroup
  - ALTOFiles: group containing all ALTO OCR files
  - ABBYYFiles: group containing all ABBYY OCR files

Repeatable: yes

content/childs: fileGrp, file

Attributes:

ID	XML ID that specifies also the type of the file group as described above	REQUIRED
USE	Determine the use of the given file group. Possible values are: Preservation, Viewing, Content	REQUIRED

The METS file-element: file

Namespace: http://www.loc.gov/METS/

Description: Every single image, ocr-xml, etc. that is part of the newspaper issue, is represented by a single file element and is assigned to its respective file group

Repeatable: yes

Content/childs: FLocat

Attributes:

ID	XML Identifier	REQUIRED
ADMID	Id reference to administrative metadata section, see amdSec	OPTIONAL
MIMETYPE	specifies the content of the file (RFC2616)	OPTIONAL
SEQ	can be used for reading order representation	OPTIONAL
CHECKSUMTYPE	Specifies the type of checksum used	OPTIONAL
CHECKSUM	Checksum of given file itself	OPTIONAL

The METS file location-Element: FLocat

Namespace: http://www.loc.gov/METS/

Description: the FLocat element is used to hyper reference the files using an URL

Repeatable: no

Content/childs: none

Attributes:

LOCTYPE	Specifies the type of reference as URL	REQUIRED
xlink:href	URL to the file	REQUIRED

The METS structure map-element: structMap

Namespace: http://www.loc.gov/METS/

Description: every ENMAP document contains a physical structure map, listing all files of the issue and can be used to store a page type map and the pagination. Beside that an ENMAP document can contain several logical structure maps. The simple ENMAP contains only the physical structure map.

Repeatable: yes

Content/childs: div

Attributes:

TYPE	Specifies the type of the structure map. physical_structmap, logical_structmap	REQUIRED
ID	XML Identifier	REQUIRED

The METS div-element: div

Namespace: http://www.loc.gov/METS/

Description:

1) In case of a physical structMap the div-elements should be used to create correlations to the single pages of the document.

2) In a logical structure map the div elements are used to build the hierarchical structure of the document. The types 'content section' and 'content unit' are used to create the structure tree nodes, where content sections are used to group together content units and can contain further content sections. Content units can contain other content units or 'content items', which are the leaves of the structure tree and are used for the physical representation. Content item values are taken from the list provided above.

Repeatable: yes

Content/childs: div, fptr

Attributes:

ID	XML Identifier	REQUIRED
DMDID	XML IdRef to the descriptive metadata section, see dmdSec	OPTIONAL
ADMID	Id reference to administrative metadata section, see amdSec	OPTIONAL
LABEL	can contain the title of a content section or a content unit	OPTIONAL
TYPE	In case of a physical structMap it can contain the page type, so far 3 types are foreseen, but that list can be extended: titlepage, contentpage, lastpage in case of a logical structMap it contains the logical type: content section, content unit, or one of the content	REQUIRED

	item types	
ORDER	Is used on top of the hierarchical order to represent the reading order	REQUIRED
ORDERLABEL	Is used in the physical structure map to represent the pagination	OPTIONAL

The METS fptr-Element: fptr

Namespace: http://www.loc.gov/METS/

Description: The fptr-element contains different possibilities to create physical references to the actual div-element/structural element

Repeatable: yes, where every new repeat equates to one derivative.

Content/childs: area, seq

The METS seq-Element: seq

Namespace: http://www.loc.gov/METS/

Description: The seq-element can be used to group two or more physical representations for one content item.

Repeatable: no

Content/childs: area

The METS area-Element: area

Namespace: http://www.loc.gov/METS/

Description: Finally the area-element is used to reference a certain area onto one page.

Repeatable: only when wrapped by a seq-Element

Content/childs: none

Attributes:

FILEID	A XML-Identifier referencing a file from the file section	REQUIRED
COORDS	A string with 4 Integer values separated by an empty space, referencing a certain rectangular area in a file by the top left and bottom right vertices	OPTIONAL
CONTENTIDS	A list of Id-references of the content file	OPTIONAL

ENMAP Examples

These and some other examples and corresponding result packages are downloadable via the Europeana Newspapers homepage. The examples can be viewed with the developed STRUCTIFY tool. The download link of the tool is available via the homepage as well. On the download page a HOWTO guideline helps to open the ENMAP examples and ENMAP deliveries from UIBK as well as from CCS produced during this project.

ENMAP Example Simple:

Example Issue: LFT_00006_BZZ_19150201

```
<?xml version="1.0" encoding="UTF-8"?>
<METS:mets TYPE="Newspaper" PROFILE="ENMAP" OBJID="LFT_00006_BZZ_19150201"
  xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/version191/mets.xsd
http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/mods.xsd http://www.loc.gov/mix/v20
http://www.loc.gov/standards/mix/mix20/mix20.xsd"
  xmlns:METS="http://www.loc.gov/METS/" xmlns:mix="http://www.loc.gov/mix/v20"
xmlns:mods="http://www.loc.gov/mods/v3" xmlns:xlink="http://www.w3.org/1999/xlink"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:mets="http://www.loc.gov/METS/">
  <mets:metsHdr CREATEDATE="2013-03-11T11:16:17" LASTMODDATE="2013-03-11T11:16:17"
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>DEA METS Engine</mets:name>
    </mets:agent>
  </mets:metsHdr>
  <mets:dmdSec ID="MODS_ISSUE_BZZ_19150201">
    <mets:mdWrap MDTYPE="MODS">
      <mets:xmlData>
        <mods:mods>
          <mods:titleInfo>
            <mods:title>Bozner Zeitung</mods:title>
          </mods:titleInfo>
          <mods:typeOfResource>text</mods:typeOfResource>
          <mods:genre>Newspaper Issue</mods:genre>
          <mods:originInfo>
            <mods:dateIssued encoding="w3cdtf" keyDate="yes">1915-02-01</mods:dateIssued>
          </mods:originInfo>
          <mods:language>
            <mods:languageTerm type="code" authority="iso639-2b">ger</mods:languageTerm>
            <mods:scriptTerm type="code" authority="iso15924">Goth</mods:scriptTerm>
          </mods:language>
          <mods:identifier type="ENP">LFT_00006</mods:identifier>
          <mods:identifier type="CatalogueIdentifier">486622</mods:identifier>
          <mods:accessCondition type="Europeana viewer">2 - Snippet view</mods:accessCondition>
          <mods:recordInfo>
            <mods:recordContentSource>LFT</mods:recordContentSource>
          </mods:recordInfo>
        </mods:mods>
      </mets:xmlData>
    </mets:mdWrap>
  </mets:dmdSec>
  <mets:amdSec ID="TECH_MD">
    <mets:techMD ID="FID-1915_02_0001-OCRMASTER-TECHMD">
      <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>3436</mix:imageWidth>
                <mix:imageHeight>4835</mix:imageHeight>
              </mix:BasicImageCharacteristics>
            </mix:BasicImageInformation>
            <mix:ImageAssessmentMetadata>
              <mix:SpatialMetrics>
                <mix:samplingFrequencyPlane>object plane</mix:samplingFrequencyPlane>
                <mix:samplingFrequencyUnit>in.</mix:samplingFrequencyUnit>
                <mix:xSamplingFrequency>
                  <mix:numerator>300</mix:numerator>
                  <mix:denominator>1</mix:denominator>
                </mix:xSamplingFrequency>
                <mix:ySamplingFrequency>
                  <mix:numerator>300</mix:numerator>
                  <mix:denominator>1</mix:denominator>
                </mix:ySamplingFrequency>
              </mix:SpatialMetrics>
              <mix:ImageColorEncoding>
                <mix:BitsPerSample>
                  <mix:bitsPerSampleValue>1</mix:bitsPerSampleValue>
                  <mix:bitsPerSampleUnit>integer</mix:bitsPerSampleUnit>
                </mix:BitsPerSample>
                <mix:samplesPerPixel>1</mix:samplesPerPixel>
              </mix:ImageColorEncoding>
            </mix:ImageAssessmentMetadata>
          </mix:mix>
        </mets:xmlData>
      </mets:mdWrap>
    </mets:techMD>
    <mets:techMD ID="FID-1915_02_0002-OCRMASTER-TECHMD">
      <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>3198</mix:imageWidth>
            <mix:imageHeight>4693</mix:imageHeight>
           </mix:BasicImageCharacteristics>
          </mix:BasicImageInformation>
          <mix:ImageAssessmentMetadata>
           <mix:SpatialMetrics>
            <mix:samplingFrequencyPlane>object plane</mix:samplingFrequencyPlane>
            <mix:samplingFrequencyUnit>in.</mix:samplingFrequencyUnit>
            <mix:xSamplingFrequency>
             <mix:numerator>300</mix:numerator>
             <mix:denominator>1</mix:denominator>
            </mix:xSamplingFrequency>
            <mix:ySamplingFrequency>
             <mix:numerator>300</mix:numerator>
             <mix:denominator>1</mix:denominator>
            </mix:ySamplingFrequency>
           </mix:SpatialMetrics>
           <mix:ImageColorEncoding>
            <mix:BitsPerSample>
             <mix:bitsPerSampleValue>1</mix:bitsPerSampleValue>
             <mix:bitsPerSampleUnit>integer</mix:bitsPerSampleUnit>
            </mix:BitsPerSample>
            <mix:samplesPerPixel>1</mix:samplesPerPixel>
           </mix:ImageColorEncoding>
          </mix:ImageAssessmentMetadata>
         </mix:mix>
        </mets:xmlData>
       </mets:mdWrap>
      </mets:techMD>

           …
     </mets:amdSec>
     <mets:fileSec>
      <mets:fileGrp ID="ImageGroup">
       <mets:fileGrp ID="OCRMasterFiles" USE="Preservation">
        <mets:file ID="FID-1915_02_0001-OCRMASTER" ADMID="FID-1915_02_0001-OCRMASTER-
    TECHMD" MIMETYPE="image/tiff" SEQ="1" CHECKSUMTYPE="MD5" CHECKSUM="2F-0F-DD-DE-
    E6-EB-4F-6B-9A-A4-08-30-E8-97-BE-A6">
         <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/1915_02_0001.tif"/>
```

```
        </mets:file>
        <mets:file ID="FID-1915_02_0002-OCRMASTER" ADMID="FID-1915_02_0002-OCRMASTER-TECHMD" MIMETYPE="image/tiff" SEQ="2" CHECKSUMTYPE="MD5" CHECKSUM="12-1A-C4-68-3B-48-E3-1C-EF-3E-54-4B-08-1E-66-3D">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/1915_02_0002.tif"/>
        </mets:file>
        <mets:file ID="FID-1915_02_0003-OCRMASTER" ADMID="FID-1915_02_0003-OCRMASTER-TECHMD" MIMETYPE="image/tiff" SEQ="3" CHECKSUMTYPE="MD5" CHECKSUM="B0-82-E2-8A-28-E6-69-17-72-4C-25-81-69-92-4C-B0">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/1915_02_0003.tif"/>
        </mets:file>
        <mets:file ID="FID-1915_02_0004-OCRMASTER" ADMID="FID-1915_02_0004-OCRMASTER-TECHMD" MIMETYPE="image/tiff" SEQ="4" CHECKSUMTYPE="MD5" CHECKSUM="10-88-E4-B8-18-47-FA-EF-16-33-62-53-C6-34-36-6A">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/1915_02_0004.tif"/>
        </mets:file>
        <mets:file ID="FID-1915_02_0005-OCRMASTER" ADMID="FID-1915_02_0005-OCRMASTER-TECHMD" MIMETYPE="image/tiff" SEQ="5" CHECKSUMTYPE="MD5" CHECKSUM="A6-EF-85-EF-C2-4E-64-B5-C8-39-04-E0-6D-F3-F7-02">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/1915_02_0005.tif"/>
        </mets:file>
        <mets:file ID="FID-1915_02_0006-OCRMASTER" ADMID="FID-1915_02_0006-OCRMASTER-TECHMD" MIMETYPE="image/tiff" SEQ="6" CHECKSUMTYPE="MD5" CHECKSUM="2A-88-BF-28-E5-88-4A-AE-32-DA-42-E9-FB-31-A0-77">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/1915_02_0006.tif"/>
        </mets:file>
        <mets:file ID="FID-1915_02_0007-OCRMASTER" ADMID="FID-1915_02_0007-OCRMASTER-TECHMD" MIMETYPE="image/tiff" SEQ="7" CHECKSUMTYPE="MD5" CHECKSUM="45-D4-47-BE-69-6C-DB-80-3F-83-92-6F-54-2A-7D-3F">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/1915_02_0007.tif"/>
        </mets:file>
        <mets:file ID="FID-1915_02_0008-OCRMASTER" ADMID="FID-1915_02_0008-OCRMASTER-TECHMD" MIMETYPE="image/tiff" SEQ="8" CHECKSUMTYPE="MD5" CHECKSUM="86-24-EE-52-7E-E1-68-10-8B-9F-0B-68-13-14-F1-97">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/1915_02_0008.tif"/>
        </mets:file>
      </mets:fileGrp>
```

```
    </mets:fileGrp>
    <mets:fileGrp ID="TextGroup">
      <mets:fileGrp ID="ALTOFiles" USE="Content">
        <mets:file ID="FID-1915_02_0001-ALTO" SEQ="1" CHECKSUMTYPE="MD5" CHECKSUM="A1-64-D9-77-DF-E7-7F-5B-7D-69-BB-15-01-5D-CF-0B">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/1915_02_0001.xml"/>
        </mets:file>
        <mets:file ID="FID-1915_02_0002-ALTO" SEQ="2" CHECKSUMTYPE="MD5" CHECKSUM="E8-10-7A-B7-0C-A6-EC-00-5D-0A-C2-2A-0E-80-27-86">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/1915_02_0002.xml"/>
        </mets:file>
        <mets:file ID="FID-1915_02_0003-ALTO" SEQ="3" CHECKSUMTYPE="MD5" CHECKSUM="06-4C-C8-AC-C5-CB-02-0C-A9-1B-CC-D0-92-EE-53-FC">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/1915_02_0003.xml"/>
        </mets:file>
        <mets:file ID="FID-1915_02_0004-ALTO" SEQ="4" CHECKSUMTYPE="MD5" CHECKSUM="09-39-F9-DF-8B-61-7C-48-76-33-3C-16-D6-82-FC-1B">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/1915_02_0004.xml"/>
        </mets:file>
        <mets:file ID="FID-1915_02_0005-ALTO" SEQ="5" CHECKSUMTYPE="MD5" CHECKSUM="43-18-81-80-C6-EC-52-00-CA-DD-5A-59-6D-61-BD-CC">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/1915_02_0005.xml"/>
        </mets:file>
        <mets:file ID="FID-1915_02_0006-ALTO" SEQ="6" CHECKSUMTYPE="MD5" CHECKSUM="A9-73-84-AA-B2-30-E6-6D-F4-A6-18-4C-E4-71-9D-FD">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/1915_02_0006.xml"/>
        </mets:file>
        <mets:file ID="FID-1915_02_0007-ALTO" SEQ="7" CHECKSUMTYPE="MD5" CHECKSUM="4F-49-22-8D-62-E1-89-31-1E-0A-84-54-47-82-55-FE">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/1915_02_0007.xml"/>
        </mets:file>
        <mets:file ID="FID-1915_02_0008-ALTO" SEQ="8" CHECKSUMTYPE="MD5" CHECKSUM="18-16-4E-D0-71-E9-9F-BE-10-6E-E0-6A-CB-AC-35-E6">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/1915_02_0008.xml"/>
        </mets:file>
      </mets:fileGrp>
    </mets:fileGrp>
```

```
  </mets:fileSec>
  <mets:structMap LABEL="Physical Structure" TYPE="PHYSICAL">
    <mets:div ID="phys0" LABEL="BZZ_19150201" TYPE="physSequence"
DMDID="MODS_ISSUE_BZZ_19150201">
      <mets:div ID="phys1" ORDER="1" ORDERLABEL="1" TYPE="page">
        <mets:fptr FILEID="FID-1915_02_0001-OCRMASTER"/>
        <mets:fptr FILEID="FID-1915_02_0001-ALTO"/>
      </mets:div>
      <mets:div ID="phys2" ORDER="2" ORDERLABEL="2" TYPE="page">
        <mets:fptr FILEID="FID-1915_02_0002-OCRMASTER"/>
        <mets:fptr FILEID="FID-1915_02_0002-ALTO"/>
      </mets:div>
      <mets:div ID="phys3" ORDER="3" ORDERLABEL="3" TYPE="page">
        <mets:fptr FILEID="FID-1915_02_0003-OCRMASTER"/>
        <mets:fptr FILEID="FID-1915_02_0003-ALTO"/>
      </mets:div>
      <mets:div ID="phys4" ORDER="4" ORDERLABEL="4" TYPE="page">
        <mets:fptr FILEID="FID-1915_02_0004-OCRMASTER"/>
        <mets:fptr FILEID="FID-1915_02_0004-ALTO"/>
      </mets:div>
      <mets:div ID="phys5" ORDER="5" ORDERLABEL="5" TYPE="page">
        <mets:fptr FILEID="FID-1915_02_0005-OCRMASTER"/>
        <mets:fptr FILEID="FID-1915_02_0005-ALTO"/>
      </mets:div>
      <mets:div ID="phys6" ORDER="6" ORDERLABEL="6" TYPE="page">
        <mets:fptr FILEID="FID-1915_02_0006-OCRMASTER"/>
        <mets:fptr FILEID="FID-1915_02_0006-ALTO"/>
      </mets:div>
      <mets:div ID="phys7" ORDER="7" ORDERLABEL="7" TYPE="page">
        <mets:fptr FILEID="FID-1915_02_0007-OCRMASTER"/>
        <mets:fptr FILEID="FID-1915_02_0007-ALTO"/>
      </mets:div>
      <mets:div ID="phys8" ORDER="8" ORDERLABEL="8" TYPE="page">
        <mets:fptr FILEID="FID-1915_02_0008-OCRMASTER"/>
        <mets:fptr FILEID="FID-1915_02_0008-ALTO"/>
      </mets:div>
    </mets:div>
  </mets:structMap>
</METS:mets>
```

ENMAP Example with structure:

Example issue: LFT_00010_LZ_19450331

```
<?xml version="1.0" encoding="UTF-8"?>
<METS:mets OBJID="#OBJID#" PROFILE="ENMAP" TYPE="unknown"
xmlns:METS="http://www.loc.gov/METS/" xmlns:mets="http://www.loc.gov/METS/"
xmlns:mix="http://www.loc.gov/mix/v20"
  xmlns:mods="http://www.loc.gov/mods/v3" xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/version191/mets.xsd
http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/mods.xsd http://www.loc.gov/mix/v20
http://www.loc.gov/standards/mix/mix20/mix20.xsd">
  <mets:metsHdr CREATEDATE="2014-02-18T12:28:21" LASTMODDATE="2014-02-18T12:28:21"
RECORDSTATUS="SUBMITTED">
    <mets:agent ROLE="CREATOR" TYPE="ORGANIZATION">
      <mets:name>UIBK</mets:name>
    </mets:agent>
    <mets:agent ROLE="CREATOR" TYPE="OTHER">
      <mets:name>run</mets:name>
    </mets:agent>
  </mets:metsHdr>
  <mets:dmdSec ID="MODS_1">
    <mets:mdWrap MDTYPE="MODS">
      <mets:xmlData>
        <mods:mods>
          <mods:titleInfo>
            <mods:title>Lienzer Zeitung</mods:title>
          </mods:titleInfo>
          <mods:identifier>LFT_00010 / 569145</mods:identifier>
          <mods:originInfo>
            <mods:dateIssued>1945-03-31</mods:dateIssued>
          </mods:originInfo>
          <mods:language>
            <mods:languageTerm>ger</mods:languageTerm>
          </mods:language>
        </mods:mods>
      </mets:xmlData>
    </mets:mdWrap>
  </mets:dmdSec>
  <mets:dmdSec ID="DMD_LS41">
    <mets:mdWrap MDTYPE="MODS">
      <mets:xmlData>
        <mods:mods>
          <mods:subject>
            <mods:topic>serial novel</mods:topic>
          </mods:subject>
          <mods:name>
            <mods:namePart>Rudolf Eichthal</mods:namePart>
            <mods:role>
              <mods:roleTerm>creator</mods:roleTerm>
            </mods:role>
          </mods:name>
          <mods:originInfo>
            <mods:place>
              <mods:placeTerm>Salzburg</mods:placeTerm>
            </mods:place>
          </mods:originInfo>
          <mods:language>
            <mods:languageTerm>ger</mods:languageTerm>
          </mods:language>
        </mods:mods>
      </mets:xmlData>
    </mets:mdWrap>
  </mets:dmdSec>
  <mets:dmdSec ID="DMD_LS46">
    <mets:mdWrap MDTYPE="MODS">
      <mets:xmlData>
        <mods:mods>
          <mods:subject>
            <mods:topic>verbatim reports</mods:topic>
```

This metadata section belongs to the content unit with DMDID=MD_LS41 in the logical structmap and was enriched with several MODS elements.

```
        </mods:subject>
        <mods:name>
          <mods:namePart>Dr. med. Kurt Oxenius</mods:namePart>
          <mods:role>
            <mods:roleTerm>creator</mods:roleTerm>
          </mods:role>
        </mods:name>
        <mods:language>
          <mods:languageTerm>ger</mods:languageTerm>
        </mods:language>
      </mods:mods>
    </mets:xmlData>
  </mets:mdWrap>
</mets:dmdSec>
<mets:dmdSec ID="DMD_LS51">
  <mets:mdWrap MDTYPE="MODS">
    <mets:xmlData>
      <mods:mods>
        <mods:subject>
          <mods:topic>poems</mods:topic>
        </mods:subject>
        <mods:name>
          <mods:namePart>Sepp Raneburger</mods:namePart>
          <mods:role>
            <mods:roleTerm>creator</mods:roleTerm>
          </mods:role>
        </mods:name>
        <mods:language>
          <mods:languageTerm>ger</mods:languageTerm>
        </mods:language>
      </mods:mods>
    </mets:xmlData>
  </mets:mdWrap>
</mets:dmdSec>
<mets:dmdSec ID="DMD_LS62">
  <mets:mdWrap MDTYPE="MODS">
    <mets:xmlData>
```

```
        <mods:mods>
          <mods:subject>
            <mods:topic>advertisement</mods:topic>
          </mods:subject>
        </mods:mods>
      </mets:xmlData>
    </mets:mdWrap>
  </mets:dmdSec>
  <mets:dmdSec ID="DMD_LS66">
    <mets:mdWrap MDTYPE="MODS">
      <mets:xmlData>
        <mods:mods>
          <mods:subject>
            <mods:topic>verbatim reports</mods:topic>
          </mods:subject>
        </mods:mods>
      </mets:xmlData>
    </mets:mdWrap>
  </mets:dmdSec>
  <mets:fileSec>
    <mets:fileGrp ID="TextGroup">
      <mets:fileGrp ID="OCRMasterFiles" USE="Preservation">
        <mets:file ID="FID-00000001-OCRMASTER" SEQ="1">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/LZ_19450331_26_01.JPG"/>
        </mets:file>
        <mets:file ID="FID-00000002-OCRMASTER" SEQ="2">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/LZ_19450331_26_02.JPG"/>
        </mets:file>
        <mets:file ID="FID-00000003-OCRMASTER" SEQ="3">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/LZ_19450331_26_03.JPG"/>
        </mets:file>
        <mets:file ID="FID-00000004-OCRMASTER" SEQ="4">
          <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/LZ_19450331_26_04.JPG"/>
        </mets:file>
      </mets:fileGrp>
    </mets:fileGrp>
    <mets:fileGrp>
  <mets:fileGrp ID="ALTOFiles" USE="Content">
   <mets:file ID="FID-00000001-ALTO" SEQ="1">
    <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/LZ_19450331_26_01.xml"/>
   </mets:file>
   <mets:file ID="FID-00000002-ALTO" SEQ="2">
    <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/LZ_19450331_26_02.xml"/>
   </mets:file>
   <mets:file ID="FID-00000003-ALTO" SEQ="3">
    <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/LZ_19450331_26_03.xml"/>
   </mets:file>
   <mets:file ID="FID-00000004-ALTO" SEQ="4">
    <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/LZ_19450331_26_04.xml"/>
   </mets:file>
  </mets:fileGrp>
 </mets:fileGrp>
</mets:fileSec>
<mets:structMap LABEL="Physical Structure" TYPE="PHYSICAL">
 <mets:div DMDID="MODS_1" ID="phys1" ORDER="1" TYPE="physSequence">
  <mets:div ID="phys2" ORDER="1" ORDERLABEL="1" TYPE="page">
   <mets:fptr FILEID="FID-00000001-OCRMASTER"/>
   <mets:fptr FILEID="FID-00000001-ALTO"/>
  </mets:div>
  <mets:div ID="phys3" ORDER="2" ORDERLABEL="2" TYPE="page">
   <mets:fptr FILEID="FID-00000002-OCRMASTER"/>
   <mets:fptr FILEID="FID-00000002-ALTO"/>
  </mets:div>
  <mets:div ID="phys4" ORDER="3" ORDERLABEL="3" TYPE="page">
   <mets:fptr FILEID="FID-00000003-OCRMASTER"/>
   <mets:fptr FILEID="FID-00000003-ALTO"/>
  </mets:div>
  <mets:div ID="phys5" ORDER="4" ORDERLABEL="4" TYPE="page">
   <mets:fptr FILEID="FID-00000004-OCRMASTER"/>
   <mets:fptr FILEID="FID-00000004-ALTO"/>
  </mets:div>
 </mets:div>
</mets:structMap>
<mets:structMap ID="logical_structmap_1" TYPE="logical_structmap">
```

```
  <mets:div DMDID="MODS_1" ID="LS1" LABEL="Lienzer Zeitung" ORDER="1" TYPE="Issue">
   <mets:div ID="LS2" ORDER="1" TYPE="content_unit">
    <mets:div ID="LS3" ORDER="1" TYPE="heading">
     <mets:fptr>
      <mets:area COORDS="238 64 2462 176" FILEID="FID-00000002-OCRMASTER"/>
     </mets:fptr>
     <mets:fptr>
      <mets:area COORDS="238 64 2462 176" FILEID="FID-00000002-ALTO"/>
     </mets:fptr>
    </mets:div>
   <mets:div ID="LS32" ORDER="3" TYPE="content_unit">
    <mets:div ID="LS33" ORDER="1" TYPE="text">
     <mets:fptr>
      <mets:area COORDS="2616 64 3432 2398" FILEID="FID-00000002-OCRMASTER"/>
     </mets:fptr>
     <mets:fptr>
      <mets:area COORDS="2616 64 3432 2398" FILEID="FID-00000002-ALTO"/>
     </mets:fptr>
    </mets:div>
   </mets:div>
   <mets:div DMDID="DMD_LS41" ID="LS41" LABEL="Das Wundermädchen von Südtirol" ORDER="6" TYPE="content_unit">
    <mets:div ID="LS42" ORDER="1" TYPE="heading">
     <mets:fptr>
      <mets:area COORDS="47 3442 869 3565" FILEID="FID-00000002-OCRMASTER"/>
     </mets:fptr>
     <mets:fptr>
      <mets:area COORDS="47 3442 869 3565" FILEID="FID-00000002-ALTO"/>
     </mets:fptr>
    </mets:div>
    <mets:div ID="LS43" ORDER="2" TYPE="copyright_note">
     <mets:fptr>
      <mets:area COORDS="47 3585 869 3664" FILEID="FID-00000002-OCRMASTER"/>
     </mets:fptr>
     <mets:fptr>
```

DMD_LS41 is the reference to the dmdSec. This content unit is composed of heading, copyright note and several text blocks. These are some of the possible content item types defined before.

```
            <mets:area COORDS="47 3585 869 3664" FILEID="FID-00000002-ALTO"/>
          </mets:fptr>
        </mets:div>
      <mets:div DMDID="DMD_LS46" ID="LS46" LABEL="Muss das Neugeborene hungern?" ORDER="7" TYPE="content_unit">
        <mets:div ID="LS47" ORDER="1" TYPE="heading">
          <mets:fptr>
            <mets:area COORDS="1184 3908 2283 4028" FILEID="FID-00000003-OCRMASTER"/>
          </mets:fptr>
          <mets:fptr>
            <mets:area COORDS="1184 3908 2283 4028" FILEID="FID-00000003-ALTO"/>
          </mets:fptr>
        </mets:div>
      <mets:div DMDID="DMD_LS62" ID="LS62" ORDER="11" TYPE="content_unit">
        <mets:div ID="LS63" ORDER="1" TYPE="text">
          <mets:fptr>
            <mets:area COORDS="27 4628 592 5266" FILEID="FID-00000004-OCRMASTER"/>
          </mets:fptr>
          <mets:fptr>
            <mets:area COORDS="27 4628 592 5266" FILEID="FID-00000004-ALTO"/>
          </mets:fptr>
        </mets:div>
        <mets:div ID="LS64" ORDER="2" TYPE="text">
          <mets:fptr>
            <mets:area COORDS="602 3996 2298 5264" FILEID="FID-00000004-OCRMASTER"/>
          </mets:fptr>
          <mets:fptr>
            <mets:area COORDS="602 3996 2298 5264" FILEID="FID-00000004-ALTO"/>
          </mets:fptr>
        </mets:div>
        <mets:div ID="LS65" ORDER="3" TYPE="text">
          <mets:fptr>
            <mets:area COORDS="2309 2791 3446 5267" FILEID="FID-00000004-OCRMASTER"/>
          </mets:fptr>
          <mets:fptr>
            <mets:area COORDS="2309 2791 3446 5267" FILEID="FID-00000004-ALTO"/>
          </mets:fptr>
        </mets:div>
      </mets:div>
      <mets:div DMDID="DMD_LS66" ID="LS66" ORDER="12" TYPE="content_unit">
        <mets:div ID="LS67" ORDER="1" TYPE="text">
          <mets:fptr>
            <mets:area COORDS="612 2795 2304 3972" FILEID="FID-00000004-OCRMASTER"/>
          </mets:fptr>
          <mets:fptr>
            <mets:area COORDS="612 2795 2304 3972" FILEID="FID-00000004-ALTO"/>
          </mets:fptr>
        </mets:div>
      </mets:div>
    </mets:div>
  </mets:structMap>
</METS:mets>
```

Example Issue: ONB_00004_18700520

```
<?xml version="1.0" encoding="UTF-8"?>
<METS:mets OBJID="#OBJID#" PROFILE="ENMAP" TYPE="unknown"
  xmlns:METS="http://www.loc.gov/METS/"
  xmlns:mets="http://www.loc.gov/METS/"
  xmlns:mix="http://www.loc.gov/mix/v20"
  xmlns:mods="http://www.loc.gov/mods/v3"
  xmlns:xlink="http://www.w3.org/1999/xlink"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/version191/mets.xsd
http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/mods.xsd http://www.loc.gov/mix/v20
http://www.loc.gov/standards/mix/mix20/mix20.xsd">
  <mets:metsHdr CREATEDATE="2014-02-18T12:28:21"
    LASTMODDATE="2014-02-18T12:28:21" RECORDSTATUS="SUBMITTED">
    <mets:agent ROLE="CREATOR" TYPE="ORGANIZATION">
      <mets:name>UIBK</mets:name>
    </mets:agent>
    <mets:agent ROLE="CREATOR" TYPE="OTHER">
      <mets:name>run</mets:name>
    </mets:agent>
  </mets:metsHdr>
```

```
<mets:dmdSec ID="MODS_1">
  <mets:mdWrap MDTYPE="MODS">
    <mets:xmlData>
      <mods:mods>
        <mods:titleInfo>
          <mods:title>Innsbrucker Nachrichten</mods:title>
        </mods:titleInfo>
        <mods:subject>
          <mods:topic>news</mods:topic>
        </mods:subject>
        <mods:originInfo>
          <mods:dateIssued>1870-05-20</mods:dateIssued>
        </mods:originInfo>
      </mods:mods>
    </mets:xmlData>
  </mets:mdWrap>
</mets:dmdSec>
<mets:dmdSec ID="DMD_LS2">
  <mets:mdWrap MDTYPE="MODS">
    <mets:xmlData>
      <mods:mods>
        <mods:subject>
          <mods:topic>news</mods:topic>
        </mods:subject>
      </mods:mods>
    </mets:xmlData>
  </mets:mdWrap>
</mets:dmdSec>
<mets:dmdSec ID="DMD_LS15">
  <mets:mdWrap MDTYPE="MODS">
    <mets:xmlData>
      <mods:mods>
        <mods:subject>
          <mods:topic>classified advertisement</mods:topic>
        </mods:subject>
      </mods:mods>
    </mets:xmlData>
  </mets:mdWrap>
</mets:dmdSec>
<mets:dmdSec ID="DMD_LS19">
  <mets:mdWrap MDTYPE="MODS">
    <mets:xmlData>
      <mods:mods>
        <mods:subject>
          <mods:topic>obituaries</mods:topic>
        </mods:subject>
      </mods:mods>
    </mets:xmlData>
  </mets:mdWrap>
</mets:dmdSec>
<mets:dmdSec ID="DMD_LS33">
  <mets:mdWrap MDTYPE="MODS">
    <mets:xmlData>
      <mods:mods>
        <mods:subject>
          <mods:topic>stock exchange</mods:topic>
        </mods:subject>
      </mods:mods>
    </mets:xmlData>
  </mets:mdWrap>
</mets:dmdSec>
<mets:dmdSec ID="DMD_LS36">
  <mets:mdWrap MDTYPE="MODS">
    <mets:xmlData>
      <mods:mods>
        <mods:subject>
          <mods:topic>imprint</mods:topic>
        </mods:subject>
      </mods:mods>
    </mets:xmlData>
  </mets:mdWrap>
</mets:dmdSec>
<mets:fileSec>
  <mets:fileGrp ID="TextGroup">
```

```
   <mets:fileGrp ID="OCRMasterFiles" USE="Preservation">
    <mets:file ID="FID-00000001-OCRMASTER" SEQ="1">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/00000001.tif"/>
    </mets:file>
    <mets:file ID="FID-00000002-OCRMASTER" SEQ="2">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/00000002.tif"/>
    </mets:file>
    <mets:file ID="FID-00000003-OCRMASTER" SEQ="3">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/00000003.tif"/>
    </mets:file>
    <mets:file ID="FID-00000004-OCRMASTER" SEQ="4">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/00000004.tif"/>
    </mets:file>
    <mets:file ID="FID-00000005-OCRMASTER" SEQ="5">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/00000005.tif"/>
    </mets:file>
    <mets:file ID="FID-00000006-OCRMASTER" SEQ="6">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/00000006.tif"/>
    </mets:file>
    <mets:file ID="FID-00000007-OCRMASTER" SEQ="7">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/00000007.tif"/>
    </mets:file>
    <mets:file ID="FID-00000008-OCRMASTER" SEQ="8">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./OCRmaster/00000008.tif"/>
    </mets:file>
   </mets:fileGrp>
  </mets:fileGrp>
  <mets:fileGrp>
   <mets:fileGrp ID="ALTOFiles" USE="Content">
    <mets:file ID="FID-00000001-ALTO" SEQ="1">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/00000001.xml"/>
    </mets:file>
    <mets:file ID="FID-00000002-ALTO" SEQ="2">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/00000002.xml"/>
    </mets:file>
    <mets:file ID="FID-00000003-ALTO" SEQ="3">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/00000003.xml"/>
    </mets:file>
    <mets:file ID="FID-00000004-ALTO" SEQ="4">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/00000004.xml"/>
    </mets:file>
    <mets:file ID="FID-00000005-ALTO" SEQ="5">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/00000005.xml"/>
    </mets:file>
    <mets:file ID="FID-00000006-ALTO" SEQ="6">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/00000006.xml"/>
    </mets:file>
    <mets:file ID="FID-00000007-ALTO" SEQ="7">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/00000007.xml"/>
    </mets:file>
    <mets:file ID="FID-00000008-ALTO" SEQ="8">
     <mets:FLocat LOCTYPE="URL" xlink:href="file:///./alto/00000008.xml"/>
    </mets:file>
   </mets:fileGrp>
  </mets:fileGrp>
 </mets:fileSec>
 <mets:structMap LABEL="Physical Structure" TYPE="PHYSICAL">
  <mets:div DMDID="MODS_1" ID="phys1" ORDER="1" TYPE="physSequence">
   <mets:div ID="phys2" ORDER="1" ORDERLABEL="1" TYPE="page">
    <mets:fptr FILEID="FID-00000001-OCRMASTER"/>
    <mets:fptr FILEID="FID-00000001-ALTO"/>
   </mets:div>
   <mets:div ID="phys3" ORDER="2" ORDERLABEL="2" TYPE="page">
    <mets:fptr FILEID="FID-00000002-OCRMASTER"/>
    <mets:fptr FILEID="FID-00000002-ALTO"/>
   </mets:div>
   <mets:div ID="phys4" ORDER="3" ORDERLABEL="3" TYPE="page">
    <mets:fptr FILEID="FID-00000003-OCRMASTER"/>
    <mets:fptr FILEID="FID-00000003-ALTO"/>
   </mets:div>
   <mets:div ID="phys5" ORDER="4" ORDERLABEL="4" TYPE="page">
    <mets:fptr FILEID="FID-00000004-OCRMASTER"/>
    <mets:fptr FILEID="FID-00000004-ALTO"/>
   </mets:div>
   <mets:div ID="phys6" ORDER="5" ORDERLABEL="5" TYPE="page">
    <mets:fptr FILEID="FID-00000005-OCRMASTER"/>
    <mets:fptr FILEID="FID-00000005-ALTO"/>
   </mets:div>
   <mets:div ID="phys7" ORDER="6" ORDERLABEL="6" TYPE="page">
    <mets:fptr FILEID="FID-00000006-OCRMASTER"/>
    <mets:fptr FILEID="FID-00000006-ALTO"/>
   </mets:div>
   <mets:div ID="phys8" ORDER="7" ORDERLABEL="7" TYPE="page">
    <mets:fptr FILEID="FID-00000007-OCRMASTER"/>
    <mets:fptr FILEID="FID-00000007-ALTO"/>
   </mets:div>
   <mets:div ID="phys9" ORDER="8" ORDERLABEL="8" TYPE="page">
    <mets:fptr FILEID="FID-00000008-OCRMASTER"/>
    <mets:fptr FILEID="FID-00000008-ALTO"/>
   </mets:div>
  </mets:div>
 </mets:structMap>
 <mets:structMap ID="logical_structmap_1" TYPE="logical_structmap">
  <mets:div DMDID="MODS_1" ID="LS1" LABEL="Zur Tagesgeschichte"
   ORDER="1" TYPE="Issue">
   <mets:div DMDID="DMD_LS2" ID="LS2" LABEL="Zur Tagesgeschichte"
    ORDER="1" TYPE="content_section">
    <mets:div ID="LS3" ORDER="1" TYPE="content_unit">
     <mets:div ID="LS4" ORDER="1" TYPE="heading">
      <mets:fptr>
       <mets:area COORDS="488 1078 1039 1141" FILEID="FID-00000001-OCRMASTER"/>
      </mets:fptr>
      <mets:fptr>
       <mets:area COORDS="488 1078 1039 1141" FILEID="FID-00000001-ALTO"/>
      </mets:fptr>
     </mets:div>
     <mets:div ID="LS5" ORDER="2" TYPE="text">
      <mets:fptr>
       <mets:area COORDS="72 1143 1442 2227" FILEID="FID-00000001-OCRMASTER"/>
      </mets:fptr>
      <mets:fptr>
```

```
        <mets:area COORDS="72 1143 1442 2227" FILEID="FID-00000001-ALTO"/>
      </mets:fptr>
    </mets:div>
    <mets:div ID="LS6" ORDER="3" TYPE="text">
      <mets:fptr>
        <mets:area COORDS="154 -4 1515 2176" FILEID="FID-00000002-OCRMASTER"/>
      </mets:fptr>
      <mets:fptr>
        <mets:area COORDS="154 -4 1515 2176" FILEID="FID-00000002-ALTO"/>
      </mets:fptr>
    </mets:div>
    <mets:div ID="LS7" ORDER="4" TYPE="text">
      <mets:fptr>
        <mets:area COORDS="73 107 1437 2228" FILEID="FID-00000003-OCRMASTER"/>
      </mets:fptr>
      <mets:fptr>
        <mets:area COORDS="73 107 1437 2228" FILEID="FID-00000003-ALTO"/>
      </mets:fptr>
    </mets:div>
    <mets:div ID="LS8" ORDER="5" TYPE="text">
      <mets:fptr>
        <mets:area COORDS="145 62 1522 1831" FILEID="FID-00000004-OCRMASTER"/>
      </mets:fptr>
      <mets:fptr>
        <mets:area COORDS="145 62 1522 1831" FILEID="FID-00000004-ALTO"/>
      </mets:fptr>
    </mets:div>
  </mets:div>
</mets:div>
<mets:div DMDID="DMD_LS9" ID="LS9"
  LABEL="Lokales und Verschiedenes" ORDER="2" TYPE="content_section">
  <mets:div ID="LS10" ORDER="1" TYPE="content_unit">
    <mets:div ID="LS11" ORDER="1" TYPE="heading">
      <mets:fptr>
        <mets:area COORDS="473 1844 1214 1905" FILEID="FID-00000004-OCRMASTER"/>
      </mets:fptr>
      <mets:fptr>
```

```
              <mets:area COORDS="473 1844 1214 1905" FILEID="FID-00000004-ALTO"/>
            </mets:fptr>
          </mets:div>
          <mets:div ID="LS12" ORDER="2" TYPE="text">
            <mets:fptr>
              <mets:area COORDS="145 1903 1522 2191" FILEID="FID-00000004-OCRMASTER"/>
            </mets:fptr>
            <mets:fptr>
              <mets:area COORDS="145 1903 1522 2191" FILEID="FID-00000004-ALTO"/>
            </mets:fptr>
          </mets:div>
          <mets:div ID="LS13" ORDER="3" TYPE="text">
            <mets:fptr>
              <mets:area COORDS="80 73 1462 2194" FILEID="FID-00000005-OCRMASTER"/>
            </mets:fptr>
            <mets:fptr>
              <mets:area COORDS="80 73 1462 2194" FILEID="FID-00000005-ALTO"/>
            </mets:fptr>
          </mets:div>
          <mets:div ID="LS14" ORDER="4" TYPE="text">
            <mets:fptr>
              <mets:area COORDS="140 85 1510 623" FILEID="FID-00000006-OCRMASTER"/>
            </mets:fptr>
            <mets:fptr>
              <mets:area COORDS="140 85 1510 623" FILEID="FID-00000006-ALTO"/>
            </mets:fptr>
          </mets:div>
        </mets:div>
      </mets:div>
      <mets:div DMDID="DMD_LS15" ID="LS15" LABEL="Verlosungen" ORDER="3"
TYPE="content_section">
        <mets:div ID="LS16" ORDER="1" TYPE="content_unit">
          <mets:div ID="LS17" ORDER="1" TYPE="heading">
            <mets:fptr>
              <mets:area COORDS="579 650 1088 702" FILEID="FID-00000006-OCRMASTER"/>
            </mets:fptr>
            <mets:fptr>
        <mets:area COORDS="579 650 1088 702" FILEID="FID-00000006-ALTO"/>
      </mets:fptr>
    </mets:div>
    <mets:div ID="LS18" ORDER="2" TYPE="text">
      <mets:fptr>
        <mets:area COORDS="145 701 1511 1208" FILEID="FID-00000006-OCRMASTER"/>
      </mets:fptr>
      <mets:fptr>
        <mets:area COORDS="145 701 1511 1208" FILEID="FID-00000006-ALTO"/>
      </mets:fptr>
    </mets:div>
  </mets:div>
</mets:div>
<mets:div DMDID="DMD_LS19" ID="LS19"
  LABEL="Verstorbene in Innsbruck" ORDER="4" TYPE="content_section">
  <mets:div ID="LS20" ORDER="1" TYPE="content_unit">
    <mets:div ID="LS21" ORDER="1" TYPE="heading">
      <mets:fptr>
        <mets:area COORDS="450 1241 1170 1304" FILEID="FID-00000006-OCRMASTER"/>
      </mets:fptr>
      <mets:fptr>
        <mets:area COORDS="450 1241 1170 1304" FILEID="FID-00000006-ALTO"/>
      </mets:fptr>
    </mets:div>
    <mets:div ID="LS22" ORDER="2" TYPE="text">
      <mets:fptr>
        <mets:area COORDS="146 1307 1505 1552" FILEID="FID-00000006-OCRMASTER"/>
      </mets:fptr>
      <mets:fptr>
        <mets:area COORDS="146 1307 1505 1552" FILEID="FID-00000006-ALTO"/>
      </mets:fptr>
    </mets:div>
  </mets:div>
</mets:div>
<mets:div DMDID="DMD_LS23" ID="LS23" ORDER="5" TYPE="content_section">
  <mets:div ID="LS24" ORDER="1" TYPE="content_unit">
    <mets:div ID="LS25" ORDER="1" TYPE="heading">
```

```
   <mets:fptr>
    <mets:area COORDS="141 1579 1482 1641" FILEID="FID-00000006-OCRMASTER"/>
   </mets:fptr>
   <mets:fptr>
    <mets:area COORDS="141 1579 1482 1641" FILEID="FID-00000006-ALTO"/>
   </mets:fptr>
  </mets:div>
  <mets:div ID="LS26" ORDER="2" TYPE="text">
   <mets:fptr>
    <mets:area COORDS="139 1639 1491 1989" FILEID="FID-00000006-OCRMASTER"/>
   </mets:fptr>
   <mets:fptr>
    <mets:area COORDS="139 1639 1491 1989" FILEID="FID-00000006-ALTO"/>
   </mets:fptr>
  </mets:div>
  <mets:div ID="LS27" ORDER="3" TYPE="text">
   <mets:fptr>
    <mets:area COORDS="135 2007 1489 2180" FILEID="FID-00000006-OCRMASTER"/>
   </mets:fptr>
   <mets:fptr>
    <mets:area COORDS="135 2007 1489 2180" FILEID="FID-00000006-ALTO"/>
   </mets:fptr>
  </mets:div>
 </mets:div>
</mets:div>
<mets:div DMDID="DMD_LS28" ID="LS28" ORDER="6" TYPE="content_section">
 <mets:div ID="LS29" ORDER="1" TYPE="content_unit">
  <mets:div ID="LS30" ORDER="1" TYPE="text">
   <mets:fptr>
    <mets:area COORDS="146 68 1399 2144" FILEID="FID-00000007-OCRMASTER"/>
   </mets:fptr>
   <mets:fptr>
    <mets:area COORDS="146 68 1399 2144" FILEID="FID-00000007-ALTO"/>
   </mets:fptr>
  </mets:div>
 </mets:div>
</mets:div>
```

```
      <mets:div DMDID="DMD_LS31" ID="LS31" ORDER="7" TYPE="content_section">
        <mets:div ID="LS32" ORDER="1" TYPE="text">
          <mets:fptr>
            <mets:area COORDS="142 75 1503 1566" FILEID="FID-00000008-OCRMASTER"/>
          </mets:fptr>
          <mets:fptr>
            <mets:area COORDS="142 75 1503 1566" FILEID="FID-00000008-ALTO"/>
          </mets:fptr>
        </mets:div>
      </mets:div>
      <mets:div DMDID="DMD_LS33" ID="LS33" ORDER="8" TYPE="content_section">
        <mets:div ID="LS34" ORDER="1" TYPE="content_unit">
          <mets:div ID="LS35" ORDER="1" TYPE="heading">
            <mets:fptr>
              <mets:area COORDS="141 1570 1494 2067" FILEID="FID-00000008-OCRMASTER"/>
            </mets:fptr>
            <mets:fptr>
              <mets:area COORDS="141 1570 1494 2067" FILEID="FID-00000008-ALTO"/>
            </mets:fptr>
          </mets:div>
        </mets:div>
      </mets:div>
      <mets:div DMDID="DMD_LS36" ID="LS36" ORDER="9" TYPE="content_unit">
        <mets:div ID="LS37" ORDER="1" TYPE="text">
          <mets:fptr>
            <mets:area COORDS="142 2083 1498 2191" FILEID="FID-00000008-OCRMASTER"/>
          </mets:fptr>
          <mets:fptr>
            <mets:area COORDS="142 2083 1498 2191" FILEID="FID-00000008-ALTO"/>
          </mets:fptr>
        </mets:div>
      </mets:div>
    </mets:div>
  </mets:structMap>
</METS:mets>
```