Nina Krstić
Museum of Naïve and Marginal Art, Jagodina, Serbia
Snežana Popović
Računarski fakultet, Belgrade, Serbia
Zoran Cvetković
Infotrend, Belgrade, Serbia

NEWSPAPER ARTICLES AS A SOURCE OF INFORMATION ABOUT ARTISTS AND THEIR WORKS

Abstract. Museums and other cultural institutions collect newspaper articles about cultural heritage from their domains. Their subjects of interest are articles published in newspapers, periodicals, magazines, etc., in the form of paper, or electronically on the Web. In this paper we would like to present our attitude towards the usability and usefulness of newspaper article contents, the method and software solutions for creating and managing the collection of so far published newspaper articles and software solutions for publishing selected newspaper items on the Museum's Web site.
Key words: hemeroteca, newspaper clips, digitalization method, semantic annotation, application

1. Introduction

In the era of computers when the everyday life spreads throughout the Web and social networks, old articles published in traditional newspaper media have become more and more unreachable and unavailable. However, it happens that some of cultural institutions for heritage protection [1] preserved old newspaper clips related to their domain in a form of classified and ordered collections, i.e. hemeroteca.

To establish a systematic collection of newspaper articles it is necessary to ensure continuity of their collecting and processing. Here, we use the term *hemeroteca*, which is the word of Spanish origin *hemeroteque* used in the meaning of collection of newspaper articles in the form of clips. [2]

A collection of articles has two aspects, the angle of experts and other users of information, and the angle of IT specialists, whose job is to enable permanent archiving and searching the contents of articles. As subject topics diverge, domain experts have to classify article information in terms of categories, usability, and trust in validity of published facts. In this context, IT specialists have to propose concepts and technologies for creating computer tools and programs for reliable and efficient archiving, as well as software solutions for effective searching and presentation of article contents.

In this paper we propose the methodology for professional validation, software processing and use of newspaper articles in order to obtain additional information and disseminate knowledge in the field of fine arts.

Appreciating the value of collected articles, Museum of Naïve and Marginal Art from Jagodina (MNMA), Serbia, decided to make them accessible and searchable. To do that, we carried out a project for digitization the newspaper article collection, hence we applied the proposed procedures and technologies of digitization processes to create, manage and control the collection of newspaper articles in MNMA.

2. Professional attitude - the approach of art historians
to the contents of paper clip collection

One of the methods for protection of cultural heritage, personality and works of a specific artist and kind of art, event or more or less significant data about the artist's domain of activity is the work on collecting, studying and systematisation of most various documents related to certain period of time in which the artist, artwork and the artistic chapter of creative individual who draws, paints, sculptures, designs etc. is born and developed, thus radiating with his/her art.

Documentation and systematisation of art is the first step in the study and final analysis of certain artistic phenomenon, a specific artist, one or more particular works, and the event significant for an individual or a group, either chronologically, thematically or factually. Depending on the inspiring motives of professionals that may develop at multiple levels, it is necessary to classify the most predominant topics, because numerous articles and published texts include multitude of various data, which should be arranged into specific segments for various reasons. The greater the number of segments, the better possibility for their presence at most diverse levels of study.

Hemeroteca can be chronologically divided in order to provide a reliable source of information about a particular period of time and the image of specific artistic phenomenon. It deals with a special event at precise moment, day, month or a year, a phenomenon, event, an artwork of a group or individual artists, activities of the institution, museum, gallery, or cultural centre where the event takes place (e.g. opening of exhibition, promotion of catalogues or any activity connected with the name, kind of art, activity and other data as key words). The news of an artist, exhibition, artwork, review of exhibition and artworks, interviews with the artist and hoe other see the artist, photographs (portraits) from ateliers, various events and reproductions, of particular artworks that represent the oeuvre; even the mention of a key word is always a precious source in studying, a necessary particle in the mosaic.

Independently of the above classification, but depending on the cause, a domain professional has to evaluate the quality of information and assess the possibility of its use. Naturally, some issues related to credibility and precision of data may emerge, such as whether the article is a valuable document, or is only a fragment of necessary or unnecessary statistics in the particular case, or the subject of study. In any case, the obtained data has the subject, i.e. kind of media that publicly made it known, the source to which we can refer, and, in case we feel any doubt, precisely determine its validity by using additional procedures. The document has its time determinant, short or more detailed data on the mere event, the artist, one or more, and may be factually supported (about the artwork(s) or the environment) by everything relevant and necessary for studying the continuity of development of one or more artists, events or the whole domain.

2.1 Data credibility and use. In order to create a file of an artist or exhibitions, the first condition is undoubtedly his/her biography or bibliography, or the list of exhibitions, editions and most various data which are any time available in systematised clip collection (hemeroteca) to the most detailed level.

It often occurs that during the revision of study, or creation of a new or amended biography or bibliography we find wrong data in a clip, where the author of the text arbitrarily relied on insufficiently valid sources, even a printing error, which is then multiplied in a whole series of further publications without the previous verification, thus becoming an inaccurate source. By comparing various segments of clip collection,

i.e. by examining the source of information (the author of text, medium that announced it, etc.), and comparing it with other available documents about the author, it is possible to determine the validity of data. The seriousness and reliability of data are more valuable if the author refers to specific source of information, which then becomes our target until the ultimate analysis and identification of precise data. The published articles become a usable source only if we can, more or less successfully, make them reliable in most cases after all data revisions, depending on the existing data in the database of an artist, event or a work. They are reliable when the circle of information is closed, i.e. when it is possible to check them.

2.2 Biographical data. Amendment of data on artists, an artist, a phenomenon, an event and particular artwork or several works is one of basic steps in the systematisation of artistic documentation.

Hemeroteca provides sources, chronology and other data precious for the study of life and work of the artist, the analysis of his oeuvre, the chronology of events in the continuity of his artistic work. All the mentioned sources of hemeroteca provide new sources for the biography of artist, dossier of organised exhibitions, and bibliography that in fact becomes a list of new precious sources that are used as a base for determination of document validity. Biographical data which are the target information often reveal multitude of additional data related to the continuity of development of life and artistic path of particular artist, his/her artistic oeuvre, even the ambience which may also be a significant secondary source for this or some other artist, phenomenon, or event that can become an independent study subject.

2.3 Bibliographical data. In the function of creation and amendment of the existing data on one or more artists, bibliography is a starting and unavoidable point at most various levels of study. We begin with hemeroteca in the case of studying any phenomenon, artist, his/her work, and even quite different entry.

In the sections mentioned in Introduction, clip collection (hemeroteca) precisely systematises data that are decisive for studying. It is a list of reliable sources whose validity is verifiable as they are published in precise time, medium or by a particular author. The continuity of development of artistic oeuvre of particular artist is not possible to perceive without precise bibliography, i.e. hemeroteca, which is to systematise bibliographical sources depending on the needs of professionals, most often chronologically, or possibly geographically, factually, etc.

2.4 Exhibitions. Presentations of exhibitions of one author or a group are a precious fact in the study of continuity of artistic phenomenon, development of artistic work, or chronology of the event, which can be a secondary source for a series of other artists, authors of exhibitions, artistic and other phenomena at certain moment, also necessary for the same or other purposes, certainly reliably provable, since the medium that published the fact, the author of source, chronological entry, probably a photograph exists as an additional witness of the event (portrait of the artist, author of exhibition, reproduction of the artwork etc.). A photograph can bring data that cannot be found in the mere text, such as who opened the exhibition, what the exhibition area in the gallery or museum looked like, who were the guests (significant or insignificant), what the concept of exhibition was, whether the exhibition was recorded, which works were on display, who the author of the photograph was, when it was made and in which format, by which camera, thus providing multitude of new sources. If the photograph is published in electronic media it may be a unique bearer of metadata, hence it is a source of new information network.

2.5 Interesting facts about the artist (trivia). Sometimes the personality of the artist and certain details of his life, in the form of popular and informal sketches amend his/her official, professional biography. Depending on the profile of a medium which publishes the news, review, essay, reproduction, a clip reveals the profile of the medium that published the text and besides, we would say, canonised necessary data and multitude, seemingly unimportant but interesting facts, thus showing the artist, his work, the event or phenomenon from the other viewpoint. These data provide us with the attractive insight into the spirit of the time, writing style of the author of information, a different approach that may reveal a new fact and new source, which denies the occurrence of unimportant information. The description of mere procedure of the ambience where the particular artwork was created, or the artist's life presented in informal way are far more interesting to the readers as well, which is also necessary for writing monographs, catalogues and other publications on one or more artists, event, phenomenon etc. This kind of interaction with the readers or audience includes seemingly trivial but interesting details that create a connection with readers, as important as all other documents which we find in well systematised hemeroteca.

3. MNMA collection of newspaper clips

MNMA has collected newspaper articles from since foundation in the 1960s. The collection contains about 35,000 newspaper clippings, arranged chronologically. The articles are marked with unique numbers. MNMA management concluded that the collected articles could be a valuable source of information about the MNMA activity and development of naïve and marginal art in Serbia and the former Yugoslavia. Also, articles sometimes contain additional facts to accompany biographies and bibliographies of artists, etc. Thus it was decided to digitize the collection of newspaper articles, expecting:

- To obtain digital copies, in order to protect original items from various types of destruction
- To structure and add content and images of items into a database, thus making possible the improvement of arbitrary search of a content
- To publish digitized newspaper articles to the museum website, in order to support the promotion of naïve and marginal artists and their artworks, etc.

Starting from the requests, we created and used overall methodology, i.e. algorithm, schematically shown in Figure 1.
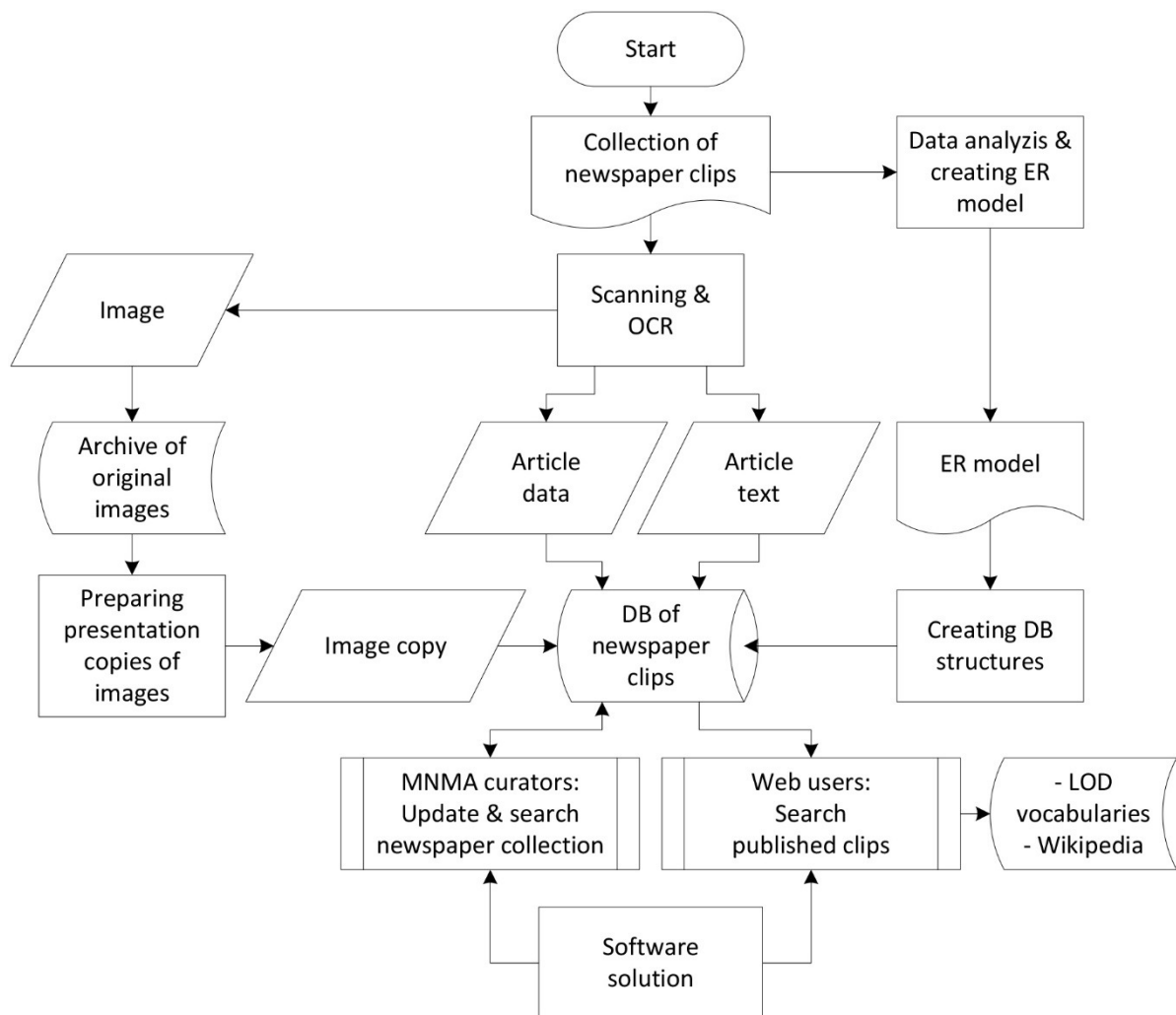
Figure 1. Process of digitizing and presenting a collection of newspaper articles

We applied the presented algorithm to digitize the entire collection of the newspaper articles at MNMA.

4. Digitizing the MNMA collection of newspaper clips

The process of digitizing the collection of MNMA newspaper clips includes the following activities:
- Preparation for scanning and scanning of newspaper clippings
- Text recognition using commercial OCR applications
- Semantic tagging of newspaper clippings
- Permanently archiving semantically-tagged images and making working copies for operational use and publishing on the Web
- Creating an application for updating and searching a database of digitized newspaper clippings
- Publishing the selected and semantically annotated newspaper clippings on the MNMA website.

4.1 Software application Hemeroteca. Scanned newspaper clips are just a large bunch of images, unless they are structured in a database and supported by software solutions. To provide the full digital protection, primarily paper-based newspaper clippings, search and effective presentation of newspaper clippings,

MNMA decided to develop the appropriate software application. The logical and physical database models are results of data analysis of news articles in details. Some of the included data are: the name of the newspaper or gazette where the article was published; title and subtitle; the text; keywords; authors, etc.

The resulting software application, Hemeroteca, is designed to support creating and updating the database, managing a digitized collection of newspaper articles and searching the database. One of typical user forms is shown in Figure 2.

Figure 2. Process of digitizing and presenting a collection of newspaper articles

4.2 Presentation of newspaper clips at the Web. The collection of newspaper articles at MNMA, among others, contains articles that can be attractive:

- When they describe less known details from the life of an artist, etc.;
- When they explain the creation of an artwork;
- When the article deals with critical review of the artist's work
- When the article is based on an authentic story, or interview with the artist, etc.

Starting from the above criteria, MNMA has decided to publish the selected articles on the museum Web site. Published images and Web pages are semantically annotated. Semantic annotation is based on Schema.org [3] and JSON-LD [4] serialization. The articles are semantically described using open vocabularies, referring to articles in Wikipedia and other relevant sources on the Web. The Figure 3 illustrates an example of a newspaper article on the MNMA Web site.

Figure 3. The example of a newspaper article on the MNMA Web site
(http://naiveart.rs/hemeroteka-image/01704v)

The part of the semantically annotated Web page source code:

```
<script type="application/ld+json" id="jsonlddata">
{
  "@context": "http://schema.org",
  "@type": "Article",
    "headline": [
      {
        "@value": "Otac Oparićke škole",
        "@language": "sr"
      },
      {
        "@value": "Father of the Oparić school",
        "@language": "en"
      }
    ],
    "description": [
      {
        "@value": "Odlomak iz knjige 'Jugoslovenski naivci', autora Milice
Maširević, upravnice Galerije samoukih umetnika u Svetozarevu.",
        "@language": "sr"
      },
      {
        "@value": "A part from the book 'Yugoslav Naive', by Milica
Maširević, director of the Gallery of Self-though Artists in Svetozarevo.",
        "@language": "en"
      }
    ],
    "keywords": [
      {
        "@value": "naivna umetnost, Oparićka umetnička grupa, škola,
naivci, samouki, Janko Brašić",
        "@language": "sr"
      },
```

```
        {
          "@value": "naive art, Oparić, artistic group, villagers, painter,
Janko Brašić",
          "@language": "en"
        }
      ],
      "inLanguage": {
        "@type": "Language",
        "name": "Serbian",
        "sameAs": "http://id.loc.gov/vocabulary/languages/srp"
      },
    "publisher": {
            "@type": "Organization",
            "name": "Novi put",
            "address": {
            "@type": "PostalAddress",
            "addressLocality": "Svetozarevo",
            "addressCountry": "Serbia"
        }
      },
    "mainEntityOfPage": {
            "@type": "WebPage",
        "name": "http://www.naiveart.rs/brasic-janko"
    },
    "image": {
            "@type": "ImageObject",
        "url": "http://www.naiveart.rs/hemeroteka-slike/01704v.jpg"
    },
      "datePublished": "1968-10-10",
      "dateModified": "1968-10-10",
      "license": "http://creativecommons.org/licenses/by-sa/3.0/deed.en",
    "author": {
        "@type": "Person",
        "name": "M. Maširević"
    },
    "about": {
            "@type": "Person",
            "name": "Janko Brašić",
            "sameAs": [
        "https://en.wikipedia.org/wiki/Janko_Brašić",
        "https://viaf.org/viaf/55544980/",
        "http://vocab.getty.edu/ulan/500195137"
            ],
            "nationality": {
                    "@type": "Country",
                    "name": "Serbia"
            }
        }
    }
  }
</script>
```

Note: The previous listing presents only a segment of the complete Web page source code and as such, it is not functional.

In this example, the content of the newspaper article is additionally semantically described using Schema.org classes (Article, Language, Organization, WebPage, ImageObject, Person, Country), and linking to LOD vocabularies [5] and corresponding sources, such as: (Library of Congress Subject Headings, Getty The Union List of Artist Names (ULAN), VIAF: The Virtual International Authority File, Wikipedia). [6]

After formal testing of the source code, search engines recognize the Web page as structured data. [7] Regarding the relevant search terms, the final result is a satisfying initial positioning in the list of search results. After certain time, it happens that the position in the search results goes down. One of the possible reasons is the stationary content, as no additional updates of this Web page are required.

5. Conclusion and future works

The collections of relevant newspaper articles, owned by the institution for the protection of cultural heritage, are a valuable source of additional information in the domain of cultural institutions. Museums of fine arts can use the collected newspaper articles as additional data to supplement biographies and bibliographies of artists, obtain the insight in the artistic creative processes and popularize visual arts in general.

In this article we proposed a method for digitizing the collection of newspaper articles, in order to protect original newspapers from damage and destruction, improve availability and make more efficient use of available information as well as contribute to the popularization of visual arts. Furthermore, we introduced the method that we developed within the project of digitalization of MNMA newspaper clip collection. The project results include software application for managing and searching the MNMA collection of newspaper clips, presentation of selected newspaper articles on the MNMA Web site, etc.

The proposed method is also applicable for digitizing collections of newspaper articles in other domains of cultural heritage such as museum objects, archaeological sites, film materials, etc.

Working on the project, we encountered some issues that required equally important, additional considerations such as handling and manipulating newspaper articles published on the Web, processing copyrights, etc. Certainly, those issues are of great importance currently, and they will be considered within the scope of further parts of our project.

Finally, we believe that collections of newspaper articles, although treated as a secondary source of data, are necessary to protect within overall digitalization and make accessible to experts and general population of Web users in the domain of culture.

References

[1]   J. Jokilehto, "*Definition of cultural heritage references to documents in history*," 2005.

[2]   "Special library." [Online]. Available: https://en.wikipedia.org/wiki/Special_library#Hemerotec. [Accessed: 29-Nov-2018].

[3]   "schema.org." [Online]. Available: https://schema.org. [Accessed: 29-Nov-2018].

[4]   M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, and N. Lindström, "*JSON-LD 1.0 A JSON-based Serialization for Linked Data*," W3C Recommendation, 2014. [Online]. Available: https://www.w3.org/TR/json-ld/.

[5]   "Linked Open Vocabularies." [Online]. Available: http://lov.okfn.org/dataset/lov. [Accessed: 29-Nov-2018].

[6]   P. Harpring, *Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works*. Getty Research Institute, 2010.

[7]   P. Mika, "O*n Schema.org and Why It Matters for the Web*," IEEE Internet Comput., 19:4(2015), 52–55.

nina.krstic@outlook.com
spopovic@raf.rs
zcvetkovic@infotrend.rs