Title Page

Multilingual research projects: Challenges for making use of standards, authority files, and character recognition

Matthias Arnold, Heidelberg Research Architecture, Heidelberg Centre for Transcultural Studies, Heidelberg University, Germany, arnold@hcts.uni-heidelberg.de

Abstract

*Articles must have the main text prefaced by an abstract of **approximately 300 words** summarising the main arguments and conclusions: a good abstract provides the reader with a complete overview of the article. This must have the heading 'Abstract' and be easily identified from the start of the main text. The abstract should also be added to the article metadata during submission.*

Academic research about digital non-Latin script (hereafter: NLS) research data can pose a number of challenges just because the material is from a region where the Latin alphabet was not used. Not all of them are easy to spot. In this paper, I introduce two use cases to demonstrate different aspects of the complex tasks that may be related to NLS material. The first use case focuses on metadata standards used to describe NLS material. Taking the VRA Core 4 XML as example, I will show where we found limitations for NLS material and how we were able to overcome them by expanding the standard. In the second use case, I look at the research data itself. Although the full text digitization of western newspapers from the 20th century usually is not problematic anymore, this is not the case for Chinese newspapers from the Republican era (1912-1949). A major obstacle here is the dense and complex layout of the pages, which prevents OCR solutions to get to the character recognition part. In our approach, we are combining different manual and computational methods, like crowdsourcing, pattern recognition, and neural networks to be able to process the material in a more efficient way. The two use cases illustrate that data standards or processing methods which are established and stable for Latin script material may not always be easily adopted to non-Latin script research data.

Keywords

*Please provide a list of **up to six keywords or phrases** that describe the subject matter of your submission, separated by semicolons. The keywords should also be added to the article metadata during submission.*

language bias; multilingual and non-Latin script research data; metadata standards; document layout analysis; optical character recognition; page segmentation;

Main Text

*The body of the submission should be structured in a logical and easy-to-follow manner. A **clear introduction section** should be provided that allows non-specialists in the subject an understanding of the publication and a background of the issue(s) involved. The remainder of the article should be divided into **appropriate subdivisions** and labelled with descriptive headers. As a rule, no more than four levels of subdivision (and subheadings) should be used. Subheadings should use sentence case. If you use more than one level of subdivision and subheading, please indicate this clearly using a style hierarchy (e.g. "Heading 1," "Heading 2," "Heading 3" in Word or LibreOffice).*

Introduction[1]

On June 25, 2019, the Centre for Asian and Transcultural Studies (CATS) at Heidelberg University opened its doors for students, teachers, and researchers. Conceptualized as collaboratorium, CATS brings together four institutions and a unique Asia library to enable dialogic perspectives on Asia and Europe (Michaels and Mittler 2019). The research collaboratorium also features a strong digital section, comprising research data in various media, formats, and scripts across Asia from both, Digital Library and Digital Humanities research sides. This forms a solid base to engage with the many challenging tasks of multilingual and non-Latin script research data. The CATS is actively involved in the ongoing process of establishing a "coordinated network of consortia tasked with providing science-driven data services to research communities" that will form a National Research Data Infrastructure.

Based at the Heidelberg Centre for Transcultural Studies (HCTS), the former Cluster of Excellence "Asia and Europe in a Global Context" and one of the four institutions that constitute the CATS, is the Heidelberg Research Architecture (HRA). This small Digital Humanities unit aims to foster and enhance digital scholarship on an institutional level and to collaborate with research projects to unfold their digital potentials (Arnold, Decker, and Volkmann 2017; Volkmann 2019). During the past decade, more than two dozen DH related research projects were realized together with the HRA, covering regions from Europe, via Ancient Egypt and the Near East, the Indian Subcontinent to East Asia. Other projects were accomplished with the Library of the Institute of Chinese Studies at the Centre for East Asian Studies, and the Library of the South Asia Institute, who's former Savifa—the Virtual Library South Asia—has become one of the two core pillars of the DFG funded Fachinformationsdienst Asien – CrossAsia. Large parts of the materials are written in non-Latin scripts (NLS), for example, in Chinese, Japanese, and Korean, or "CJK" (The Unicode Consortium 2020, 712). Thanks to a good co-operation with the University Library, it was possible for the institute libraries to co-design, for example, multilingual and NLS aspects the central library catalog. As a result, users can now search a number of data fields in original script in HEIDI, the Catalogue for libraries of Heidelberg University, as well as in the Union Catalog of the South West German Library Consortium (Südwestdeutscher Bibliotheksverbund, SWB). The new version of the object and multimedia database of

[1] This article provides a selection of references to web pages via the OpenDACHS citation repository (Arnold, Lecher, and Vogt 2020). OpenDACHS comprises services and workflows for researchers which allow the archiving of cited resources, the generation of DOI identifiers which can be used in research publications, and the creation of library catalog records for each cited resource. Links provided by OpenDACHS point to a landing page which contains both URL's, the original website, and the archived copy. This enables the research community to retrace the references and verify the cited argumentation even at a time, when the original content of the web site has changed, or the website itself no longer exists (OpenDACHS - Centre for Asian and Transcultural Studies 2020).

Heidelberg University, heidICON, now includes a number of features that allow multilingual metadata to be entered in dedicated fields, based on a close collaboration of University Library and CATS staff. Many of the new functionalities were missing from the database system and had to be programmed and implemented by the database creators.

Although these examples show that much can be achieved when domain experts from area studies experienced with the handling of NLS data are included in the shaping of new functionalities in individual systems, it is clear that many challenges remain. As I will describe in the article, material in NLS comprises many specific challenges when one takes a closer look at individual aspects of the data.

To make digital material sustainable, machine-readable, and re-usable, international standards for metadata and encoding need to be applied. As impressively visualized in (Riley 2010) for just the cultural heritage sector, many different metadata standards exist. One important industrial standard has to be mentioned here, that is The Unicode standard. It defines individual characters from different scripts:

> The Unicode Standard is the universal character encoding standard for written characters and text. It defines a consistent way of encoding multilingual text that enables the exchange of text data internationally and creates the foundation for global software. (The Unicode Consortium 2020, 1)

Since its first release in 1991, the standard quickly gained importance and is as indispensable as it is a basic core of any digital research today. Nevertheless, not everyone may know about the many issues involved with it, especially when it comes to NLS material. For example, the encoding of individual code points and discussion about variant characters, or the provision of fonts to display the code points are prominent subjects in the discussions. Another aspect is the management of the Unicode standard itself, which is related to the bias in academic DH work towards the English language, or the inclusion of certain cursive scripts like hieratic in the standard (Fiormonte 2012; Asef and Wagner 2019; Gülden, Krause, and Verhoeven 2020).

For the following sections, I selected two use cases from existing HRA projects that will illustrate two different aspects of the complex tasks related to NLS material.

In the first use case I will discuss examples where we were able to expand a metadata standard—in this case VRA Core 4 XML—to adopt to specific NLS features. If one, for example, needs to filter metadata records by the specific transliteration schema applied, or if one just wants to output dates in their original script, the information has to be entered into the record in the first place. I will argue that this means the data standard used has to be aware of such kind of information and allow it in the data schema. In our case, as I will describe below, we developed an extension of the VRA Core schema to be able to include a way to specify the applied transliteration system, and to add dates in original script and their respective writing format.

In the second use case I will introduce some of the challenges we are facing on the way to produce full text from Republican era Chinese newspapers. As we shall see, the sheer complexity of the layout, together with the variety of its content, and CJK specific features, prevent automatic OCR procedures from being successful. “Brute force” approaches like typing or double-keying full runs of periodicals are possible, but very costly. The approach I introduce below splits the processing into separate steps. I will argue that the segmentation tasks were our first main showstopper, so we made a number of experiments, like engaging the crowd for specific tasks. I will further show that applying machine learning algorithms

can be helpful to define the internal structures of pages, like columns or registers. Our most promising approach towards segmentation is the use of neural network models trained on ground truth for the page segmentation. Combining the results of these approaches can, as we shall see, pave the way towards the partially automated creation of full text.

The use cases represent only two examples of the many challenges that may be hidden under the surface of NLS material. It is my hope that our approaches to tackle NLS material will be helpful to colleagues who conceptualize their own NLS projects to get a clearer picture of what kind of workflows they may need and which possible pitfalls they may encounter. I hope the article can also be of interest to digital librarians or DH coordinators to provide better advise to interested researchers. Raising the awareness of certain limitations of existing metadata standards and software solutions within and beyond the community of researchers working with multi-lingual and NLS material is an important pre-requisite to solving these issues. Solutions for this kind of data are not rocket science, and many issues can be overcome with sharing experiences and a decent collaborative effort. I hope this article inspires the re-use of our approaches and possibly readers will provide feedback about their ways to solve issues with NLS material.

Use case 1: Standards[2]

In the cultural heritage community, metadata schemata for documenting non-textual objects such as LIDO (Coburn et al. 2010) and VRA Core 4.0 ("VRA Core - a Data Standard for the Description of Works of Visual Culture" 2018) are becoming increasingly important. Meta-aggregators like the German Digital Library (DDB) and Europeana allow the public to discover works from the comfort of their screens, while research infrastructures like the Digital Research Infrastructure for the Arts and Humanities (DARIAH), the Common Language Resources and Technology Infrastructure (CLARIN), and the CLARIAH mergers, together with their national counterparts—in Germany—DARIAH-DE, CLARIN-D, or in the future the CLARIAH-DE research infrastructure and the successful consortia of the National Research Data Infrastructure (NFDI) initiative, provide the working ecosystems to researchers and their research data. The metadata schemata are envisioned to be flexible enough to record the diverse collections of Europe. However, if we wish to include collections and research data from areas of study that do not deal exclusively with objects from the western cultural sphere, they need to be capable to cope with different types of information, lest they be omitted. That is, careful consideration must be given to the metadata schemata themselves, and how they shape our understandings through the availability and correct labeling of data.

The HRA's intensive engagement with NLS research projects and our objective to produce interoperable and sustainable data led us to international metadata and encoding standards (Riley 2010). Requests of researchers who wanted to ask their data questions such as how to sort records chronologically based on eras or reign periods (chin. 年號 *nian hao*, jap. 年号 *nengô,* originally a motto chosen by the reigning monarch), or how to filter by transliterations according to a certain schema, or how to produce a list of agents who wrote inscriptions during a certain period in time, directed us deeper into the application and implementation of data standards.

[2] Parts of this chapter were presented by former HRA member Agnes Dober in the poster "Transcultural Metadata: An exploration of the way our metadata is culturally limited" at the DARIAH-DE Grand Tour 2018.

One of the HRA tools is Ziziphus, a web-based editor for descriptive image metadata in VRA Core 4 XML. Its form-based interface allows researchers to edit VRA XML metadata without having to learn and write XML. Using an XML metadata standard makes it easier to share the data and have it re-used. We chose VRA Core 4 because it is specifically designed to encode metadata for visual media, and for its clear distinction between metadata describing the physical work and metadata describing the images that depict it.

For texts written in, for example, Japanese, it is quite natural to assume that their digital versions will include the original script, sometimes a translation, and metadata transcribed in Latin script. This is perhaps less obvious for descriptive metadata about visual resources, especially of works created in regions that mostly do not use the Latin alphabet. But let us look at an example:

> 葛飾北斎、「神奈川沖浪裏」、天保 2-4 年、多色刷木版画。

Readers of Japanese (and probably also of Chinese) will understand what this is about: 葛飾北斎 is the name of the artist Katsushika Hokusai, 「神奈川沖浪裏」 is the title of the art work "*Kanagawa oki nami ura*" which translates to "Under the Wave off Kanagawa", 天保 2-4 年 is the dating to the years "Tenpō 2-4" which can be converted to "ca. 1829–1833", and 多色刷木版画 (*tashokuzuri mokuhanga*) describes the technology used as "polychrome woodblock print". The Japanese text makes use of traditional Chinese characters, or 漢字 *kanji*, but the characters can also be written using the syllabic Hiragana (平仮名 or ひらがな) script, which renders artist name and work title like this:

> かつしか ほくさい、かながわおきなみうら。

For readers of English, the example can be translated and expressed in Latin script as:

> Katsushika Hokusai, "Under the Wave off Kanagawa", ca. 1829–1833, polychrome woodblock print.

One can see a number of problematic areas in this example, regarding language, date, and authority. A) The relation between Japanese and English record is not obvious, since they do not seem to share much at first sight, not even the same digits in the date. They need to be connected in a way that makes clear, for example, what is the original title of the work, in which script is it written, and possibly how should it be pronounced. B) The date is recorded in two different notations, in a Gregorian and a Japanese format. That means, if for example, machines (or databases) are asked to sort Japanese work records chronologically, the original date needs to be machine readable (so that a machine understands the relevant characters as a date), or the Japanese date needs to be converted into a (western) date format. C) Entities like artist name, work title, and work technique are recorded in different languages and different scripts. To make clear, which name or term/concept is meant and to prevent disambiguation issues, they should be linked to authoritative lists of names, work titles, and techniques, respectively. Ideally, these authority files contain the entity or term in different languages, like English, or Japanese.

This example illustrates some of the challenges we met while we developed our VRA Core metadata editor Ziziphus. Although this metadata standard was very flexible for describing

the art work, in a number of cases we had to expand it to accommodate features rarely seen in "western" art works but rather common for NLS metadata[3]. We discussed our ideas with members of the VRA Data Standards Committee and documented our suggestions in an open access file (Arnold 2013). We developed expanded schema files, "vra-strictCluster.xsd" and "vraCluster.xsd", and published them on GitHub. The following examples make use of data from the CCO sample records, where Hokusai's "Great Wave" is also included. I'm using them to discuss some of the specific encoding problems we faced and to demonstrate how our VRA Core 4 extension can help solving them. On a larger scale I aim at raising awareness of a bias in individual metadata standards based on the prominent use of Latin script metadata examples and the perhaps inadequate role of NLS material examples, which may even lead to the loss of parts of the original information, and how a solution may be approached.

The guide book "Cataloging Cultural Objects - a Guide to Describing Cultural Works and Their Images"(CCO) (Baca et al. 2006) is a "manual for describing, documenting, and cataloging cultural works and their visual surrogates. The primary focus of CCO is art and architecture, including but not limited to paintings, sculpture, prints, manuscripts, photographs, built works, installations, and other visual media. CCO also covers many other types of cultural works, including archaeological sites, artifacts, and functional objects from the realm of material culture" (Visual Resources Association Foundation 2019a), and is further supported by the online examples on CCO Commons (Visual Resources Association Foundation 2006). It aims at providing standards for *data content*, illustrating the many ways of combining standardized *data structures*, especially *Categories for the Description of Works of Art* (CDWA) and *VRA Core, Version 4.0*, with standards for *data values*, such as authority vocabularies like the Getty Thesauri. (Visual Resources Association Foundation 2006, xi f.)

A project completed by the Heidelberg Research Architecture in 2018 used the CCO Commons examples to provide a full set of XML records encoded in the VRA Core 4.0 metadata schema. Our intention was to help other users comprehend the XML examples by providing valid XML records, and to publish the XML records on the CCO website itself, on the Heidelberg University's research data repository HeiDATA, and on GitHub. Besides, the project was an important use case for testing our VRA Core 4 metadata editor Ziziphus.

CCO Commons provided more than 100 examples to illustrate how metadata should be recorded in a best possible way. They were organized into ten categories, with an additional, but unfinished section "Examples from the Print Publication" (Visual Resources Association Foundation 2019b). Unfortunately, the examples from CCO Commons were just published as texts in HTML format. On the VRA Core Support Pages, a set of 41 example records organized into nine categories "are provided to demonstrate how the VRA Core structure can be applied to various combinations of Work, Image and Collection in several different categories." They are not identical to the ones on CCO Commons, though there is some overlap. These samples provide data also in XML, for example no. 43, „Japanese handscroll".

To illustrate our arguments in this article, we chose "Category Examples: 7 Prints and Drawings", "Example 51: Print in a series" from the CCO Commons set as reference. This is one of the famous woodblock prints by Hokusai (Fig. 1).

[3] (Dushay 2013) provides good background information regarding the issues related to working with and processing of CJK scripts in a pragmatic way, for an introduction about the different script systems and their encoding in the Unicode standard see (The Unicode Consortium 2020, chap. 6 "Writing Systems and Punctuation" and 18 "East Asia").

Figure 1: 葛飾北斎 Katsushika Hokusai (1760–1849), 神奈川沖浪裏 "Under the Wave off Kanagawa" (Kanagawa oki nami ura), also known as "The Great Wave", from the series 富嶽三十六景 "Thirty-six Views of Mount Fuji (Fugaku sanjū rokkei)", ca. 1830–32. Polychrome woodblock print, ink and color on paper, The Metropolitan Museum of Art, JP1847 (license: CC0 1.0).

Scripts and transliteration

A typical metadata record with NLS has to distinguish between the data in original script, their transliteration to Latin script following a specific set of rules, and often a set of translations. In our example, the original title 神奈川沖浪裏 was transliterated following the Hepburn schema to "Kanagawa oki nami ura" and translated to English as "Under the Wave off Kanagawa". In this case, the work record also has a "popular" title "The Great Wave", and belongs to the series 富嶽三十六景 (*Fugaku sanjūrokkei*, "Thirty-six Views of Mount Fuji").

```
<titleSet>
   <title pref="true" lang="jpn" script="Jpan" type="creator">神奈川沖浪裏
     </title>
   <title pref="false" lang="jpn" script="Latn"
     transliteration="hepburn" type="other">Kanagawa oki nami
     ura</title>
   <title pref="false" lang="eng" script="Latn" type="translated">Under
     the Wave off Kanagawa</title>
   <title pref="false" lang="eng" script="Latn" type="popular">The Great
     Wave</title>
</titleSet>
```

A number of schemata utilize an `xml:lang` attribute for recording data about the language of a record described, with the TEI P5 standard being a prominent example. Our experiences showed that this to some degree limits its use for recording data in languages not using the Latin script.

One point is that `xml:lang` allows the recording of script, but does not support transliteration. There is, of course, a way to use extension subtags or private use subtags (Philipps and Davis 2009, chap. 2.2.6 and 2.2.7). But in an attribute `xml:lang="ja-Latn-x-hepburn"` the subtag only introduces some "hepburn" value within a private use subtag. Since the private use subtags allow almost any content, the substring "hepburn" could mean anything. The use of Latin script for Japanese data does not imply if it is the original, a translation, a transliteration, or something else. This is why we suggest to make the attributes explicit. An element containing the attributes `lang="jpn" script="Latn" transliteration="hepburn"` clearly states that the data is Japanese, written in Latin script, and that it is using the "hepburn" transliteration. One may argue that differentiation of transliteration schemas is not a major feature, and for most East Asian languages the everyday use of transliterations is quite low, indeed[4]. However, the picture changes completely when other regions with other script systems are taken into consideration, like the Ancient Egyptian scripts (Davies and Laboury 2020), or texts in Indian languages, for example the resources on the GRETIL platform ("GRETIL - Göttingen Register of Electronic Texts in Indian Languages" 2020), which provide all digital texts in transliteration. A further argument is that `xml:lang` conflates information of different kinds into one node. This has an impact to the processability of the data, because one has to parse the many combinations of language, script, and the private use subtags that may contain additional information like the transliteration schema.

Another argument in our case was interoperability with MODS XML records. In 2010, the guidelines developed in the Digital Library Federation Aquifer Initiative were combined with the MODS User Guidelines. DLF Aquifer was very clear about `xml:lang`:

> "The use of the `xml:lang` attribute is not recommended. The use of the `script` and `transliteration` attributes is recommended if applicable" (DLF Aquifer Metadata Working Group 2009, 100).

[4] An exception may be library cataloging, where author names and titles are transliterated. In Germany, the German National Library wrote a handbook "Practical guidelines for CJK records," which also defines transliteration standards: for Chinese ISO 7098 (Pinyin), for Japanese DIN 32708 (Hepburn), for Korean the "Revised Romanization rules 2000" (AG Kooperative Verbundanwendungen der AG der Verbundsysteme 2017).

Today, the MODS XML standard allows the use of both, `xml:lang`, and a combination of language, script and transliteration (Library of Congress and MODS Editorial Committee 2018).

As a side note, from the terminology of the MODS standard one can see its being deeply rooted in Latin-script data or even letter-based scripts. In a personal communication with Professor Michael Radich, chair of Buddhist Studies at my institute, he rightfully noted that "transliteration" in the context of NLS metadata can itself be problematic. Strictly speaking, he argues, the use of the term "transliteration" with a script that does not have "letters", but instead, has e.g. single glyphs that represent whole syllables (Japanese *kana,* Indic scripts), or is ideographic-pictographic (Chinese, hieroglyphs etc.) is not really correct. Instead, it would be better to speak of "transcription". He further notes: "There may then be a difficulty, then, deciding whether to apply this principle. It is clear that in talking e.g. about representation of Sanskrit in Chinese (**菩薩**, *pusa*, Bodhisattva) or English in Mandarin (**薩克斯風**, *sakesifeng*, Saxophone), it is best not to talk of "transliteration", but rather, a "transcription", because the resulting word is not in a script that has letters. I think that for transcription into a script without letters, the word "transliteration" is clearly not suitable; but for transcription out of such a script, into e.g. Roman, I think it is OK."

Based on these arguments and intensive discussions with experts from the field, we decided to implement the individual attributes, but kept the possibility of generating `xml:lang` tags for data exchange.

In our VRA Core 4 extension we are using the following attributes:

- `@lang`, following ISO 639-2B, *Codes for the Representation of Names of Languages*: Alpha-3 codes arranged alphabetically by the "bibliographic" English name of language, which, for example, uses "chi" for "Chinese" instead of "zho" for "Zhongwen", and is a good in-between of 639-1 and 639-3.
- `@script`, following ISO 15924, *Codes for the Representation of Names of Scripts*
- `@transliteration`, encoding the romanisation scheme, based on ALA-LC guidelines

Alternative calendrical dates

The VRA schema currently expects the encoding of dates according to the Gregorian calendar, and the element description requests ISO 8601 as data standard (VRA Core Data Standards Committee 2007, chap. Date). In 2019, the updated ISO 8601-2:2019 was published. It now contains a number of extensions, including uncertain dates and sections from the Extended Date/Time Format (EDTF) Specification. However, recording dates in formats different from the Gregorian calendar is quite common in NLS regions, for example era names (like Tenpō 天保, or Chunxi 淳熙), local calendars of ethnic group or nationalities (like Nepal Sambat), or calendars specific to religions (like the Hebrew Anno Mundi notation).

```
<dateSet>
   <date type="creation">
      <earliestDate>
         <date circa="true" type="about">1830</date>
         <alternativeNotation>文政 13・天保元年 (Bunsei 13 / Tenpō 1)
         </alternativeNotation>
      </earliestDate>
      <latestDate>
         <date circa="true" type="about">1832</date>
         <alternativeNotation>天保３年 (Tenpō 3)</alternativeNotation>
    </latestDate>
   </date>
</dateSet>
```

In our extension, we kept the `<date>` tag for the notation of a Gregorian calendar date, and added the element `<alternativeNotation>` to record the date as string in its original script. It is also possible to specify language and script here. For our use cases this sufficed, although, for example, reign names could also be transformed into controlled vocabulary lists.

Inscriptions

We noticed that data from inscriptions can vary a lot, which brought up several problems in encoding. Some of the issues we encountered are that the creator of the inscription can be different from the creator of the art work itself, or in some cases, multiple agents may have added multiple inscriptions to a single work. In addition, inscriptions can have specific calligraphic styles, signatures may be handwritten, but also be a seal of the agent. The issues with language, script, and transliteration also apply here.

```
<inscriptionSet>
   <inscription>
      <author refid="69033717" vocab="viaf">Katsushika, Hokusai</author>
      <position>top left</position>
      <text lang="jpn" script="Jpan" type="signature">前北斎為一筆</text>
      <text lang="jpn" script="Latn" transliteration="hepburn"
          type="signature">Zen Hokusai Iitsu hitsu</text>
      <text lang="eng" script="Latn" type="translation">From the brush
          of Iitsu, formerly Hokusai</text>
   </inscription>
</inscriptionSet>
```

As a partial solution, we added an `<author>` element to `<inscription>`, and also allowed `@lang`, `@script`, and `@transliteration` for the `<text>` subelement (see above). We decided that the encoding of stylistic elements, like the calligraphic style of an inscription or a signature, were beyond the level of detail we needed. This information could go into the description. In the future, it would also make sense to add the main language information for an individual inscription in the parent `<inscription>` element and inherit the value to the sub-elements where they can be changed if needed. This would remove some of the current redundancy in the attribute values.

Wrap-up

When working with NLS material, issues with metadata standards occur not as rarely as one might hope. Our expansions of the VRA Core 4 schema are pragmatic and straightforward.

Other modifications of existing schemata may be more challenging, for example, discussions of issues related to Unicode, as mentioned above.

The reflection of, for example, the bias in academic Digital Humanities publications towards the English language has been discussed before (Fiormonte 2016; Mahony 2018).

> Even if our spoken words are not in English, the computer systems that we rely on, with their ones and zeros, respond to and are dominated by the American Standard for Information Exchange (the ASCII code); they display browser pages encoded in HTML with their US-English defined element sets; and transfer data marked-up in the ubiquitous XML with its preference for non-accented characters, scripts that travel across the screen from left to right, and the English language-based TEI guidelines. English is very much the language of the Internet and has become the *lingua franca* of the web. (Mahony and Gao 2019, sec. 3)

On an international DH level there are the ADHO Special Interest Group GO::DH (Global Outlook::Digital Humanities) and the ADHO Standing Committee on Multi-Lingualism and Multi-Culturalism Committee (MLMC) to help the DH community to "become more linguistically and culturally inclusive in general terms, and especially in the areas where linguistic and cultural matters play a role".

The release of version 4 was a major milestone in the development of VRA Core, because it elevated the standard to an XML schema hosted by the Library of Congress. It also opened the door to the systematic usage of authoritative data (using, for example, `@vocab` and `@refid`) while keeping string-based data in the standard using the `<display>` element. Although NLS data was not the focus of its development, it was made flexible enough to allow our extensions. With new developments towards semantic data models, the VRA-RDF-Project (Mixter [2014] 2017) and it's VRA Ontology help sustain VRA Core data and open the way for data re-use within future knowledge systems.

Use case 2: Digitization and full text

Research projects are often driven towards areas that are understudied, neglected, or were not academically studied at all. In many cases, and especially for material written in non-Latin scripts, this often implies that research material, or research data has to be collected, digitized, and compiled in the first place. Only then can researchers apply systematical work or start to use algorithmic approaches to the content of the material. However, collecting data to answer a distinct set of research questions in a structured way can be a challenge on its own, if the material only exists in analog form. Often the sources are dispersed, locked or hidden away, exist in different incomplete partially overlapping sets, or are poorly preserved. It is important to acknowledge that the collection of such material itself forms an integral part of research data, especially if it is provided in digital format and—ideally—open access.

One of the projects that devoted itself to the collection of such material is the "Early Chinese Periodicals Online" (ECPO) project. Originally a by-product of a larger research project (Hockx, Judge, and Mittler 2018; Sung, Sun, and Arnold 2014), it developed into a larger platform where more than 300 mostly Republican era publications can now be read online in open access (Arnold and Hessel 2020, sec. Introduction).

Predecessors of this project started with research questions that could not be answered without a compilation of the resources. Therefore, we collected and digitized the data and created first databases where we structured the material, and made it readable online, and searchable via metadata. As of today, an impressive number of over 300.000 scans are available open access, and users can search for textual content through our manually created

analytical records in the form of bibliographical metadata and subject headings. In addition, ECPO provides structured information about the publishing history of each publication, which usually includes notes on frequency, format and size, price, publisher and prominent agents, and holdings in selected libraries. For the entertainment newspapers included in ECPO, the project added a summary of content.

An automatic generation of this kind of detailed and informed information is currently not possible. We therefore strongly believe in the importance of this manual editing. I add two more reasons: Firstly, the systematic compilation of metadata on the publication level is very hard to get at any other place. For instance, we developed an ingest workflow to establish a page sequence for each publication. This implies checking dates and issue numbers (which both may be unreliable), but also recording changes in format or price. In addition, we also take notes on possible gaps in the holdings of a number of institutions, or record the periods a publication was suspended, sometimes pointing to the respective editorials. Even when full text versions of all publications may eventually be available, this kind of information will still be hard to extract. Secondly, we produce bilingual metadata. It is our hope, that the combination of Chinese and English metadata may help making these important sources more accessible to non-readers of Chinese as well.

Creating these records is, however, very time-consuming and expensive. To further improve the usability of our database and to make a significant step towards readability of Chinese newspapers more generally, we have started to look closer on ways to transform our image scans into machine-readable text.

Chinese Republican era newspapers

Producing searchable full text for early print media can be a challenge even with western language material ("EuropeanaTech Insight - Issue 13: OCR" 2019). The German OCR-D initiative started in 2015 to coordinate the developments in OCR for printed historical texts with a focus on German language prints from the 16th to 19th centuries. It continues the efforts of the EUC project IMPACT and can be seen as parallel approach to the current EUC project READ, which focuses on handwritten text recognition of archival records.

Chinese characters are very complex, and significantly different from Latin based script systems (Magner 1974; Wilkinson 2017, chap. 2: Script and Calligraphy). With all language specific variants, the 26 Latin letters form a group of about 500, while there are over 50.000 Chinese characters alone, without taking the many variants into account. The current version 13 of the Unicode standard (March 2020) adds another 4939 CJK ideographs in the "CJK Unified Ideographs Extension G" to the total number of 93858 defined CJK Unified Ideographs which are easily accessible in the Unihan database.

Figure 2: Jing bao 晶報 (The Crystal), April 21, 1939, pages 1+4. A typical complex page layout. In ECPO: https://kjc-sv034.kjc.uni-heidelberg.de/ecpo/publications.php?magid=1&isid=20&ispage=1

In addition, the document layout of Chinese publications can be quite complex, with entertainment newspapers at the extreme side (Fig. 2). The narrow print set and very dense layout, where headlines often stretch across multiple columns, or where multi-part text blocks may be printed without any optical separators, cannot be processed by OCR engines at the moment. Furthermore, typical pages feature calligraphic headers and stretches of advertisements, adding even more to the layout complexity.

Until very recently, therefore, OCR-ing the early Chinese printing press has been considered almost impossible. The only way to produce reliable full text for a long time was double-keying—extremely labor- and cost-intensive: a quote by a double-keying agency for the newspaper *You xi bao* 游戲報 "Entertainment", a small weekly running 1897 to 1908 (562 issues, ca. 3000 pages) was about 20.000 USD for producing the raw text (3USD per 1000 characters). Adopting this quotation to a newspaper like *Jing bao* 晶報 "The Crystal" (1919-1940) would result in approximately 240.000 USD (almost 10.000 pages, more than four times the number of characters per page). Only few commercial endeavors have been undertaken in this direction, even within China, for example the new edition of late 19th ct. Shanghai daily *Shenbao* 申报 "The Shun Pao" (1872-1949), produced by double-keying specialist Greenapple Changsha (in Germany accessible for researchers via the DFG funded Fachinformationsdienst Asien – CrossAsia.

A number of Libraries in China started digitization projects of newspapers including those from the Republican era (Xiao 2017; Li and Li 2016). Many of these resources are providing

“full text“ (*quanwen* 全文) (Xiao, Wu, and Zeng 2015; Zhang 2019). However, this usually means that the full pages can be read from image scans, and only author-title data with some subject heading can be searched. For example the prestigious "Republic of China Periodical Full-text Database" 民国期刊全文数据库, which is part of the huge digital “Index to Chinese Newspapers and Periodicals” (*Quanguo baokan suoyin* 全国报刊索引) only offers image scans in “full text” mode. Only a very small number of newspapers which are marked with “OCR” provide “real” full text, for example *Xin Shidai* 新时代, and *Dongfang zazhi* 东方杂志. In other cases, the full text may have been produced, but was not yet made accessible to the public (Fang 2019.6; Zhang 2019). Consequently, the improvement of in-depth processing beyond the digitization as image scans is one of the key suggestions, Li and Li make in their study, besides improving the selection of material, including of archival material, and opening the process more to digital publishing the results (Li and Li 2016, 42f). In some projects, author-title indexing was outsourced to specialized agencies, which often employed junior high school students and high school students to analyze pages and do the typing. A review of the outcomes and suggestions to improve it was subject in studies like (Xiao 2017; Xiao and Huai 2017; Zhang 2019). In a recent report on the status of reprinting and publishing of Republican newspapers, Fang Zijin calls for more attention to tabloids (*xiao bao* 小報), to go beyond the reproduction of image scans towards their in-depth exploration and systematic organization (Fang 2019.6, 31).

Other commercial databases have started to provide full text to their clients, for example the Global Press Archive from East View Information Services, Inc., or the *Quanguo baokan suoyin* 全国报刊索引 from Shanghai Library. The biggest non-commercial program is the full-text digitization of the complete Buddhist canon by the Chinese Buddhist Electronic Text Association (CBETA, 中華電子佛典協會) in Taiwan, which took more than a decade to complete and was largely carried out by volunteers from the worldwide Buddhist community.

Only very few western scholars are working on methods to improve text recognition, an exception is the Chinese Text Project, which ran its own OCR experiments (Sturgeon 2018; 2020). It is, however, not so trivial to adapt recognition of characters from woodblock prints to the modern newspaper prints. It is even harder to adopt results from recognition of handwritten texts, even if these also contain Chinese characters. Therefore, the impressive outcome of the Kuzushiji Recognition challenges with the Kaggle Jupyter Notebook environment cannot be re-used so easily. The problems with OCR and multi-lingual data have been recognized and in 2018, the Northeastern University published the report “A Research Agenda for Historical and Multilingual Optical Character Recognition,” where they make nine recommendations “for improving historical and multilingual OCR over the next five to ten years” (Smith and Cordell 2018).

Although OCR engines significantly improved in recent years, they still do not produce useful results with Chinese texts from before the 1950s, whereas commercial double-keying agencies may have separate (higher) pricing for material dating to even “before 1990.”

Within the ECPO project, we chose *Jing bao* 晶報 “The Crystal” as data set for testing. *Jingbao* was published from March 3, 1919 to May 23, 1940 in Shanghai. It first appeared every third day, and later switched to daily. It started as insert to the *Shenzhou* (China) *Daily* 神州日報 but became independent afterwards. The newspaper employed a large number of

authors, including famous writers like Bao Tianxiao 包天笑 (1876-1973) and Zhou Shoujuan 周瘦鵑 (1895-1968), and also included cartoonists like Ding Song 丁悚 (1891-1969), Huang Wennong 黃文農 (1903-1934), and Zhang Yuguang 張聿光 (1885-1968). After its independence from Shenzhou Daily in the 1920s, *Jing bao* changed its editorial style and began to publish texts criticizing government and administration. With its sharp, satirical texts, *Jing bao's* audience grew steadily and the newspaper became known on a national level. Through guest editors and a "literary café" it reached for that time impressive circulation of more than 50.000 issues. This was the highest number compared to the other contemporary tabloids, the so-called "little papers" (*xiao bao* 小報). Together with *Luo Bin Han* 羅賓漢 "Robin Hood", *Fu er mo si* 福爾摩斯 "Holmes", and *Jin gang zuan* 金剛鑽 "The Diamond", *Jing bao* was counted one of the so-called "Four Heavenly Kings" (*si da jin gang* 四大金剛). As other tabloids, *Jing bao* addresses a broad readership and features diverse types of text, from news reports to poems or serialized literature, and also includes visuals, from photograph to painting to cartoon. Of this newspaper we manually analyzed more than 14.000 items (5368 articles, 3001 images, 5849 advertisements), mostly from April issues, and also manually produced full text for about 5000 advertisements.

Challenges with OCR

On our way towards the generation of full text we ran a number of experiments with *Jingbao*. In one experiment we tested a number of OCR systems in processing full page scans. We ran the tests in 2017 and included Abbyy Finereader v.12, Transkribus from v.03, Larex, and Abbyy Cloud OCR SDK. As a result, we had to state that none of the systems we tested was able to produce anything useful from the full pages, with recognition rates below 10%. For western-language texts, the German Research Foundation (DFG) states that "results with less than 99.5% accuracy are inadequate for manual transcription". They also require that "for printed works dating from 1850 or later, the full text must be generated in addition to image digitization," however, without mentioning NLS materials. In terms of OCR software and its usability they do not give "any conclusive recommendations" (Deutsche Forschungsgemeinschaft 2016, chap. 3.4 "Full text generation"). Instead, they are referring to the OCR-D project, which has begun to publish its own guidelines and documentations.

Figure 3: Segment from Jing bao 晶報 (The Crystal), April 18, 1919, page 2, detail. Full passage with emphasis characters.

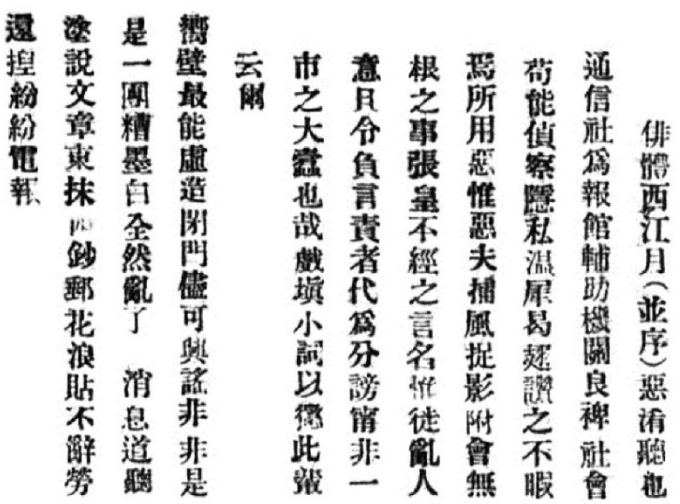

Figure 4: Segment from Jing bao 晶報 (The Crystal), April 18, 1919, page 2. After image clean-up.

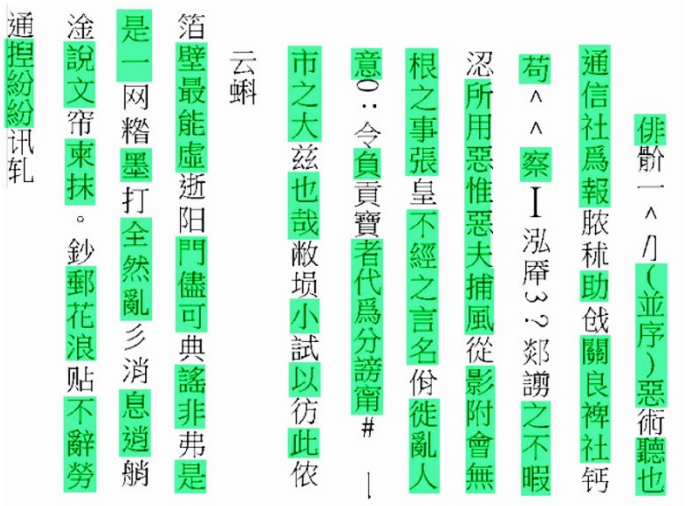

Figure 5: Segment from Jing bao 晶報 (The Crystal), April 18, 1919, page 2. OCR result, the characters marked in green were correctly recognized (ca. 63%).

In a next step, we manually produced smaller segments from articles and processed these with Abbyy Cloud OCR SDK. This was more promising, but brought another problem to the light: in the early 1920s many texts feature special characters for emphasis (Fig. 3). These additional markers were used extensively, and surprisingly often each character within a passage had an emphasis character at its side. The presence of emphasis characters also led the OCR processing to fail. However, this changed after applied a number of manual image pre-processing steps: we cropped individual segments from the page, enhanced the raster image by removing noise and intensifying contrast, and manually removed the emphasis characters. With such an improved image Abbyy Cloud OCR SDK was able to recognize more than 60 percent of the characters out-of-the-box (Fig. 4-5).

Although a CER shortly below 40% is not really a desirable result, it showed that processing individual text segments is a promising way to proceed. With an additional step in image pre-processing it is possible to remove special markers next to the characters, like emphasis markers. The results of the experiment also made clear that we needed good document segmentation algorithms before we can process the scans. Unfortunately, it turned out that automatic segmentation of complex pages like the Chinese entertainment newspapers is still a

big challenge, even with western newspapers. For example, the "ICDAR Competition on Recognition of Documents with Complex Layouts" (RDCL2019) presented material from contemporary magazines and technical articles, and these layouts do not compare to the complexity of the Chinese newspapers.

Engaging crowds

Within ECPO we therefore focused on segmentation. At that time, we got in contact with Pallas Ludens, a Heidelberg start-up specializing in crowdsourcing solutions. They had just issued a call for projects to get into contact with real world academic projects that might be interested in testing their services. Luckily, our material was exciting for them so they soon agreed on a pilot project with us. In a number of preparatory talks, we formulated criteria for the optimal outcome, where we harmonized our annotation requirements with their professional experiences on a feasible workflow with their user interfaces. One crucial point in defining the tasks for the crowd was to atomize these tasks and not to mix different levels of annotation. For example, we originally wanted to get annotations of the page structure, e.g. background and actual page, or marginalia, masthead, print area, and fold. Later we dropped this more structural annotations to have the crowd focus on the individual content boxes. Besides, with the content boxes available and correct labels assigned, the page structure becomes implicit in the outcome, and thus a separate annotation step was not necessary. Eventually, the request was to draw a box around everything that looked like a visual unit, based on a number of examples we provided, and give each box a label from a small pre-defined list.

The pilot was performed by a small and closely supervised crowd, who manually identified information blocks on the scans, drew bounding boxes, and assigned labels to them. They used an instance of the Pallas Ludens user interface, which was developed with a strong focus on work-efficiency and usability, tailored to our specific use case. Identification of the blocks was surprisingly good, and was further improved after some feedback rounds. However, since the crowd was unable to read Chinese they were not able to identify semantic units, i.e. decide, which boxes belong to the same group, e.g. an advertisement, or an article. We therefore assigned the task of grouping individual boxes into meaningful semantic units to a reader of Chinese. While this might have become a very efficient workflow for us, things turned out differently. Pallas Ludens was so good, they were acquired by a larger company and had to stop all external co-operation—including the one with us.

Nevertheless, the outcome of the pilot taught us crowdsourcing our material is indeed possible, even with a crowd who does not understand the language. It can produce very good results, especially if participants can read Chinese, are supervised, and if user interfaces with excellent usability are provided. The collaboration with an external start-up, even if short-termed due to the circumstances, proved to be very successful. While the project provided them a "fancy" use case with a number of challenges to their workflows, their professional approach towards finding a solution for us, their fresh "outsider's" look on our material and their high motivation gave the project a very positive stimulus and became a great inspiration. As a result, we completed layout analysis (bounding boxes and corrections) for four years of *Jing bao* (1919-1922), altogether 927 page-scans, and the semantic grouping of bounding boxes for all issues of April 1919. Together with our partner eXist Solutions, we launched a new tool for annotating and semantic grouping in Summer 2020, where we implemented a number of usability features based on our experiences with Pallas Ludens. With this tool we can store annotations with coordinates and labels in web annotation format in an XML database, where we can process them further.

Algorithms looking for patterns

In another approach we looked closer at the elements that visually structure the pages ("Binnenstrukturen"). The human eye can easily spot registers and text segments, supported by the various separators and different font types and sizes of the characters. In fact, the type area does have a grid structure, for example 12 horizontal registers containing eight characters of vertical text. (Fig. 6) But these registers are not so strictly followed: headlines and texts may cross multiple registers; some articles can stretch three 8-character registers but appear in two 12-character registers. To readers of Chinese, with their contextual knowledge, even the sequence of the individual segments can be determined easily. The task is quite different for the computer, since it only "sees" individual pixels. Looking closer at the separators, they turn out to be rather diverse: lines may not be straight or continuous, they may in fact be waved lines, consist of linked dots or squares, or even be just a sequence of symbols. In addition, the noise from digitized microfilm, marks from folding the newspaper, or parts where cuts appear or the paper was torn, all add to the complexity. (Fig 7-8)

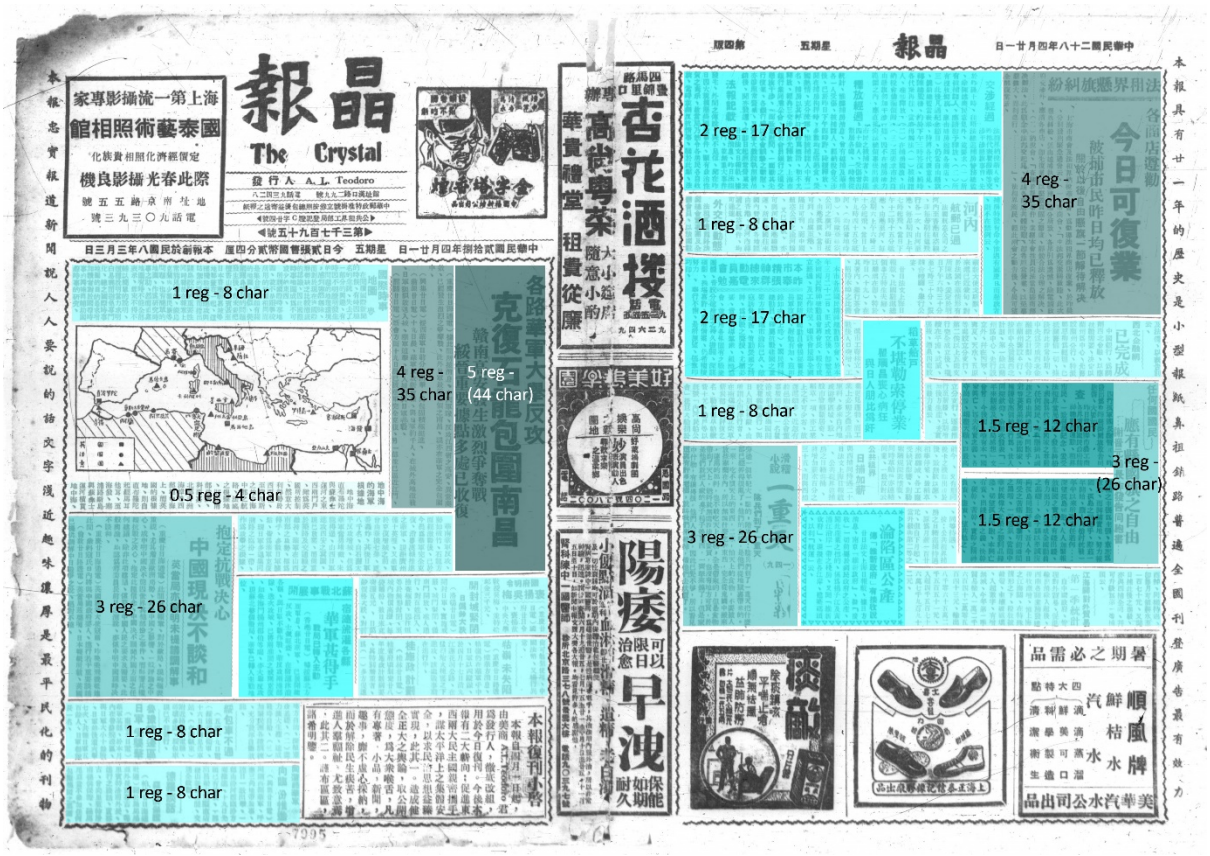

Figure 6: Jing bao 晶報 (The Crystal), April 21, 1939, pages 1+4. Visualization of horizontal registers ("columns") and their respective number of characters per line.


稻草船戶
不堪勒索停業
罷順昌喪心病狂
與日人朋比爲奸

Figure 7: Jing bao 晶報 (The Crystal), April 21, 1939, page 4, detail. Separating elements (1)


滑稽小說
一重天
陰兵們到了一重天
（一四九）
日捕加薪
公共租界
淪陷區
傳「維新政府」

Figure 8: Jing bao 晶報 (The Crystal), April 21, 1939, page 4, detail. Separating elements (2)

Looking into the text itself, another issue typical in East Asian material can be observed: the reading direction varies. Some headlines are written horizontally, left to right, especially when western letters occur, for example in the masthead. However, horizontal text written left-to-right is an exception. The typical reading direction in these materials is vertical and right-to-left. This often also applies to the article headings. However, some headlines or sub-headings may be written horizontally and right-to-left. (Fig 9) Within text passages of type “article” these structuring features are usually easy to discern. This is different for the sections of type “advertisement”. Here, basically anything may happen, horizontal and vertical, left-to-right and right-to-left texts can all be seen within a single advertisement. In addition, diagonal or curved text lines, as well as special characters, rare fonts, or inverted text may be found here. The varying reading directions of text especially within the advertisements resulted in a very high workload for double-keying, which took about 10 hours per double-page scan (“fold”).

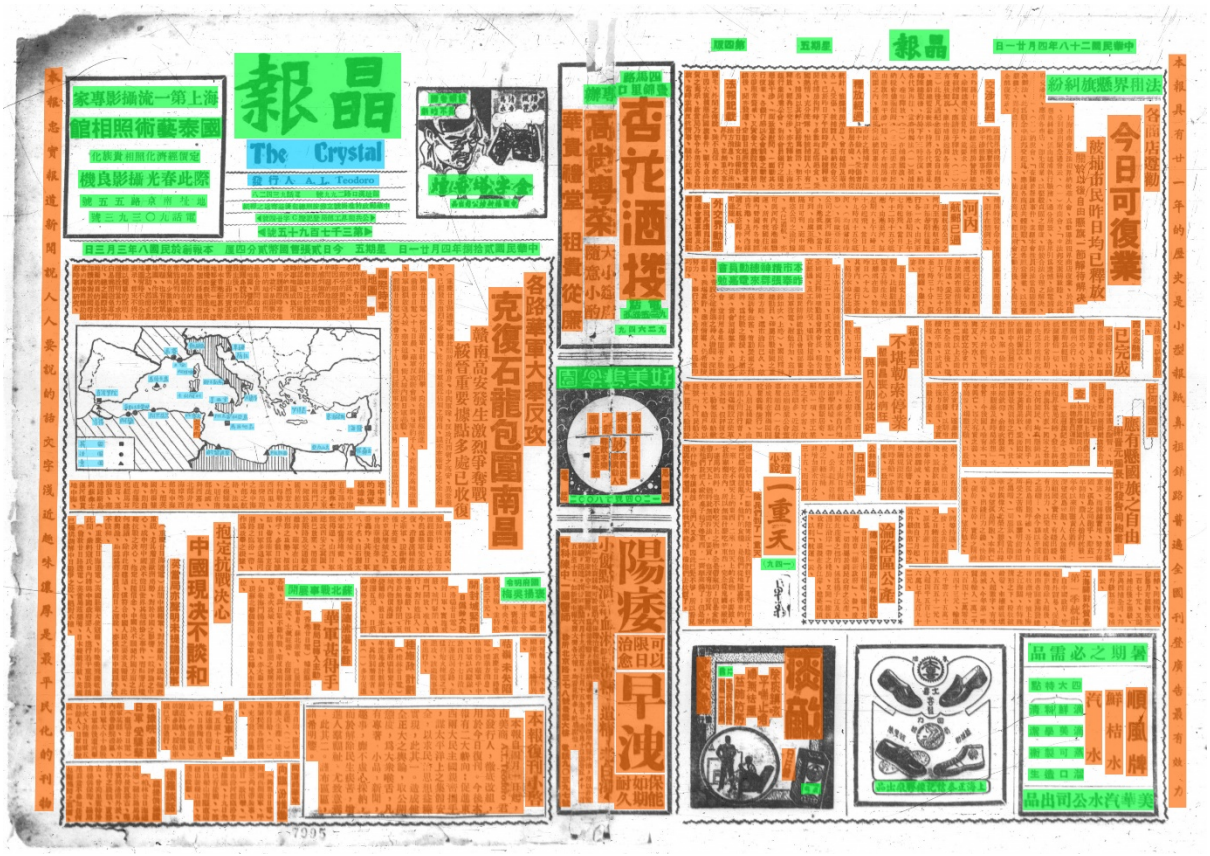

Figure 9: Jing bao 晶報 (The Crystal), April 21, 1939, pages 1+4. Reading directions: orange = vertical right-left, green = horizontal right-left, blue = horizontal left-right

Quite recently, the computational approach to line detection in historical documents has made significant progress (Grüning et al. 2019). The project currently runs experiments of the adaptation of line segmentation together with Prof. Stiehl's Image Processing Research Group at the Department of Informatics, Hamburg University.

Utilizing neural networks

With the advent of neural networks and deep learning algorithms in Digital Humanities within the last years, automatically processing of newspapers from page scans to full text has become a realistic vision. In fall 2019, the project has begun to implement "dhSegment - A generic deep-learning framework for Historical Document Processing" (Ares Oliveira, Seguin, and Kaplan 2018) and developed a prototype for training the model on ECPO data.

From the experiments with the prototype it became clear that we will need to apply different approaches to different tasks. We use the bold lines on the page to detect the printing area and thereby also the marginalia. We detect the headers with the large characters in one step, and images in another. We use the typical strong frames around advertisements for detection. (Fig. 10) This provides us with another way to find the actual texts: by deducting the areas from the page which are not defined as text areas (header, marginalia, advertisements, images), the parts which are text remain. Thus, we are able to focus on the actual text areas, which mostly correspond to items of type "article", in the next processing steps.

Figure 10: Jing bao 晶報 (The Crystal), April 21, 1939, pages 1+4. Separators highlighted: green = advertisement, dark blue = image, light blue = separating elements.

Our basis is the outcome of a short-term project funded by the Field of Focus III, Heidelberg University in fall 2019, where we created two small sets of ground truth: page segmentation with bounding boxes and labels, and full text with semantic units and reading direction (blind double-keying). The data was published on GitHub (Fig. 11). We also aim to help in the further development of these methods by training models for specific non-Latin script features, like the changing reading directions, or vertical text.

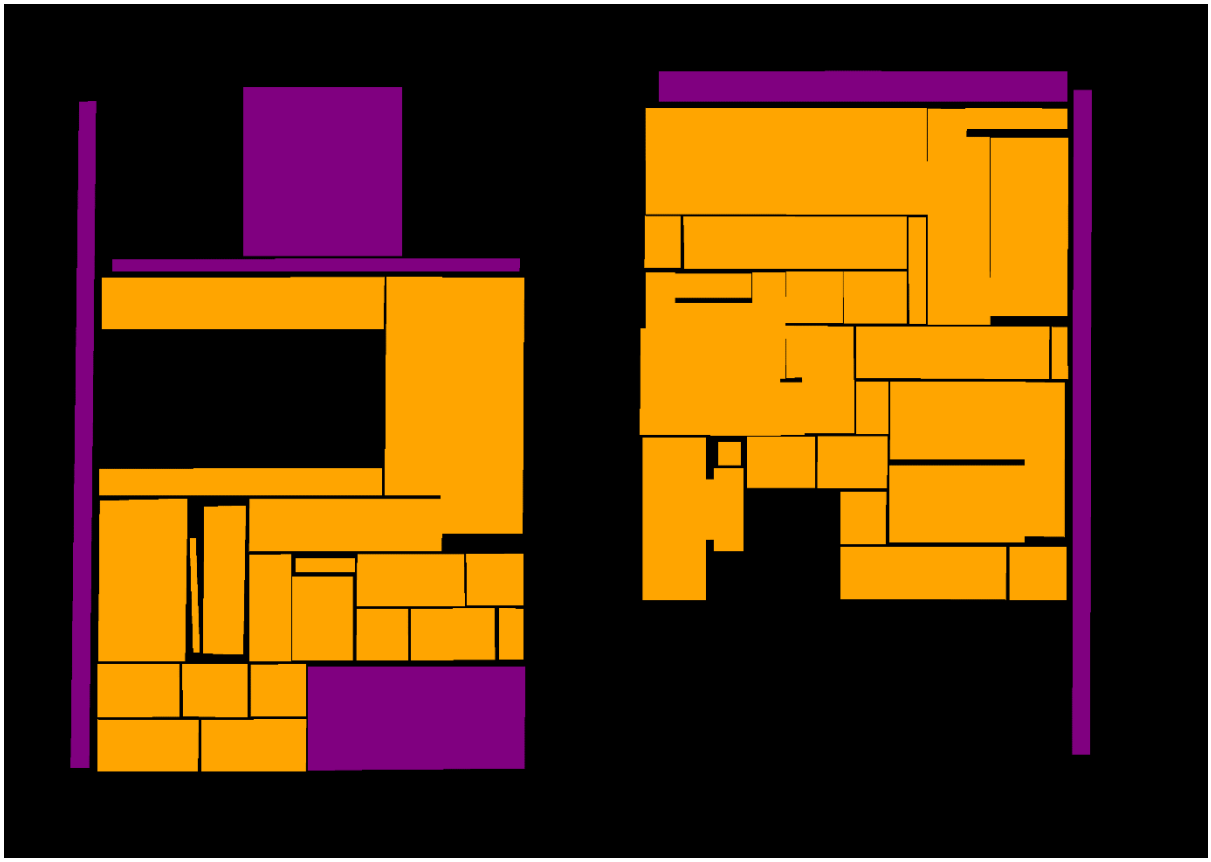

Figure 11: Jing bao 晶報 (The Crystal), April 21, 1939, pages 1+4. Detection of content types: advertisements and images ignored, purple = header and marginalia, orange = text (i.e. "articles").

Wrap-up

Not all issues discussed in this chapter are specific to Chinese newspapers of the Republican era. Many of them are not even specific to non-western script material. In fact, the language or script printed on a scan does not matter to most of the image (pre-)processing algorithms. Still, they reveal challenges in the material that current algorithms are not always capable to cope with. Often the reason is simply that solutions were developed with the focus on one distinct group of material, for example medieval manuscripts, or 18th and 19th century German language and Latin script newspapers. Adapting these algorithms to different sets of material sometimes is not trivial.

This is one of the reasons why the project is actively engaged in the "Working group newspapers and magazines" of the Digital Humanities in the German speaking regions association (DHd). There, the author presented the project at the workshop „OCR - Herausforderungen und Lösungen für Zeitungen & Zeitschriften" ("OCR – challenges and solutions for newspapers and magazines"), November 11, 2019, at the University Library of Frankfurt/Main University (presentations). We hope that the close collaboration with experts from the field can raise the awareness of challenges still awaiting solutions within non-Latin script materials. For us, broadening our own horizon is very important, we can always learn from experiences that projects with "non-NLS" material have made and how they approached and solved various problems, and possibly adopt solutions to NLS material.

We do not expect a solution that solves all these problems with republican era Chinese newspapers in the near future. Some processing steps may still require manual work. After segmentation the process is not finished, there may be other challenges with line and

character detection, also with character recognition and improving the recognition processes with special dictionaries and possibly crowd-based text improvements. When segmentation is working and text recognition producing decent results, this does not mean all problems are solved. Quite the contrary will be the case: we can then proceed into a number of new directions. One will be to test out how these workflows can be adopted to older material - texts written with Chinese characters go back to even before the first millennium BCE. This will show how scalable or flexible the solutions are with material using the same script, for example the variations in print layout, different text medium, and different writing styles. A logical follow-up question will be concerned with adopting the processes to other scripts, but there is still a long way to go. Another direction will be related to questions about what to do with the texts that will then be available. This would mean to open the Chinese texts to the core of what Digital Humanities related to text is currently working on based on Latin script material. For digital Chinese studies, this will become a new research field.

Conclusion

In this article I discussed two aspects of the challenges that research projects using digital resources with NLS pose. To make digital material sustainable, machine-readable, and re-usable, international standards for metadata and encoding must be applied. Taking VRA Core 4 as example, I showed that there may be features in research data that cannot easily be encoded in these standards. To be able to filter records by the transliteration schema applied, or to just output dates in their original script, the information has to be entered into the record, which means the data standard has to be aware of the information and allow it in the data schema. We therefore developed an expansion of the VRA Core schema that includes a way to specify the applied transliteration system, and offers a place to add dates in original script and their respective writing format.

In the other use case I discuss the challenges we are facing on the way to produce full text from Republican era Chinese newspapers. The sheer complexity of the layout, together with the variety of its content and CJK specific features prevent automatically processing from being successful. “Brute force” approaches like double-keying are possible, but very costly. The approach I introduced offers other alternatives, like engaging the crowd for specific tasks which do not even require the knowledge of the language. The most promising approach is to enhance segmentation using models trained on ground truth. I present the first results of this process, which are very promising.

There are other challenges with NLS material. To allow machines to process data in a semantically meaningful way, entities in the texts need to be found, identified, and related to authority files. There are many different authority files available, on agent names, geographic places, individual works, or concepts and terms, for example the Getty Vocabularies, the Integrated Authority File (GND), or the Virtual International Authority File (VIAF). These are usually managed by professional organizations, in the given samples the Getty Research Institute, the German National Library, and the Online Computer Library Center (OCLC), respectively. In recent years community driven authorities have become increasingly important, like DBpedia, which provides the structured data from Wikipedia pages in a way that it allows it use for the semantic data systems, Wikidata, a multilingual knowledge base that brings multilingual structured data from Wikimedia projects together, or regional alternatives to Wikipedia encyclopedias, like the Chinese 百度百科 Baidu Baike. Projects intending to share their data in machine readable format can profit a lot by linking their named entities to these authorities. In the example given above, the author of the inscription

葛飾北斎 Katsushika Hokusai is referenced to the record in VIAF: `<author refid="69033717" vocab="viaf">Katsushika, Hokusai</author>.`

However, there has to be a bi-directional exchange. Projects should not only enhance their own data by pulling information, in many cases the authority files could also benefit significantly from domain specific vocabularies. In the ECPO project we recorded many agent names in a separate module, the Agents service (Arnold and Hessel 2020, chap. 2: "The Agents service"). This is a very important module, since the sources only give names. The different names need to be assigned to individual persons, or "agents," which may sound trivial, but the disambiguation of names can easily become a challenging research task. For example, we recorded 335 occurrences of *bianzhe* 編者 (editor), or 125 of *bianjishi* 編輯室 (editorial office), where it may be very difficult to identify the actual person behind the role. While in these cases the ambiguity may remain, in other cases our project can provide some improvements. The Chinese newspapers included many names of western people, be it politicians, movie stars, or other celebrities. Very often the western names were not printed in Latin letters, but their names transliterated using Chinese characters. Since at that time the transliteration into Chinese was not regulated, the Chinese names can vary, in terms of pronunciation (the syllables chosen) and character (which of the characters with that pronunciation was used). For example, we identified twenty-five different Chinese names for the actress Constance Bennett, e.g. 蓓耐康司登 Beinai Kangsideng, 裴配康斯登 Peipei Kangsideng, 裴萊脫康絲登 Peilaituo Kangsideng, 彭乃脫康司登 Pengnaituo Kangsideng. Looking into the international authority files, we found that VIAF only includes one Chinese name 康斯坦斯·班尼特 Kangsitansi Bannite, which is a modern rendering of the name taken from Wikidata, while the GND does not list any name variant.

This not only illustrates a gap on the multi-lingual side of authority files, which needs to be filled. It also shows yet another instance where NLS data is still under-represented. On the other hand, it makes clearly visible what huge potential specialized academic databases have to contribute their data to the public. For the ECPO project we decided to prepare variant names for submitting them to the GND. Other projects should also consider if they want to share their data.

Sharing experiences, data and workflows with larger communities or the public is always an important facet in the outreach of digital projects. For projects working with or on NLS material this is also true. Perhaps even more so, because, as I have shown, the challenges are hidden in the details. Data standards or workflows that are working very well may stumble upon NLS material or reveal hidden pitfalls. When new projects are conceptualizing data templates, it is a good practice to let them provide data samples – the typical ones, but also—and even more importantly—the exceptions. The more of the possible exceptions one knows in advance, the better can a data structure be, and the more robust will the system be. For the work on NLS material we need to still learn more about these exceptions.

Acknowledgements

The author is grateful for the insightful and helpful comments of (in alphabetical order): Janice Eklund, Joan Judge, Xiaoli Ma, Bridget Madden, Duncan Paterson, Michael Radich, Cosima Wagner, and Susan Jane Williams on an earlier version. Many thanks also to the projects research assistant Xie Jia, who helped compiling and assessing the Chinese secondary sources.

Competing Interest

The author has no competing interests to declare.

List of Figures:

Works cited

All references cited within the submission must be formatted according to the ***Chicago Manual of Style 16th Edition, author-year system*** *and listed at the end of the main text file.* Authors are strongly encouraged to use a citation manager to manage and format their references *(e.g. Zotero, Paperpile, Mendeley). The journal reserves the right to charge a fee to authors whose references require significant copy-editing, formatting, and/or research during production. Please see below for sample formats.*

AG Kooperative Verbundanwendungen der AG der Verbundsysteme. 2017. “Praxisregeln zur CJK-Erfassung (Aktualisierung).” https://doi.org/10.25354/nls2020.08.36.

Ares Oliveira, Sofia, Benoit Seguin, and Frederic Kaplan. 2018. “DhSegment: A Generic Deep-Learning Approach for Document Segmentation.” In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 7–12. IEEE. https://doi.org/10.1109/ICFHR-2018.2018.00011.

Arnold, Matthias. 2013. “Suggestions for Possible Extensions of VRA Core 4 (Draft).” https://doi.org/10.25354/nls2020.08.37.

Arnold, Matthias, Eric Decker, and Armin Volkmann. 2017. “Digital Humanities Strategies in Transcultural Studies.” Working paper. Heidelberg. https://doi.org/DOI:10.11588/heidok.00023729.

Arnold, Matthias, and Lena Hessel. 2020. “Transforming Data Silos into Knowledge: Early Chinese Periodicals Online (ECPO).” In *Heuveline, Vincent , Gebhart, Fabian Und Mohammadianbisheh, Nina (Hrsg.): E-Science-Tage 2019: Data to Knowledge*, 95–109. Heidelberg: heiBOOKS. https://doi.org/10.11588/heibooks.598.c8420.

Arnold, Matthias, Hanno Lecher, and Sebastian Vogt. 2020. “OpenDACHS: Ein Citation Repository zur nachhaltigen Archivierung zitierter Online-Quellen.” In *Heuveline, Vincent , Gebhart, Fabian und Mohammadianbisheh, Nina (Hrsg.): E-Science-Tage 2019: Data to Knowledge*, 236–37. Heidelberg: heiBOOKS. https://doi.org/10.13140/RG.2.2.23094.34885.

Asef, Esther, and Cosima Wagner. 2019. “Workshop Report „Non-Latin Scripts in Multilingual Environments: Research Data and Digital Humanities in Area Studies” – Biblioblog.” *Biblioblog* (blog). January 18, 2019. https://doi.org/10.25354/nls2020.08.38.

Baca, Murtha, Patricia Harpring, Elisa Lanzi, Linda McRae, and Ann Whiteside. 2006. *Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images*. Chicago: American Library Association. https://vraweb.org/resources/cataloging-cultural-objects/.

Coburn, Erin, Richard Light, Gordon McKenna, Regine Stein, and Axel Vitzthum. 2010. “LIDO v1.0 - Lightweight Information Describing Objects.” ICOM CIDOC. https://doi.org/10.25354/nls2020.08.39.

Davies, Vanessa, and Dimitri Laboury, eds. 2020. *The Oxford Handbook of Egyptian Epigraphy and Paleography*. *The Oxford Handbook of Egyptian Epigraphy and*

Paleography. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190604653.001.0001.

Deutsche Forschungsgemeinschaft. 2016. “DFG Practical Guidelines on Digitisation.” Deutsche Forschungsgemeinschaft (DFG). https://doi.org/10.25354/nls2020.08.40.

DLF Aquifer Metadata Working Group. 2009. “Digital Library Federation / Aquifer Implementation Guidelines for Shareable MODS Records.” https://doi.org/10.25354/nls2020.08.41.

Dushay, Naomi. 2013. “CJK with Solr for Libraries, Part 1.” *Discovery Grindstone* (blog). October 29, 2013. https://doi.org/10.25354/nls2020.08.42.

“EuropeanaTech Insight - Issue 13: OCR.” 2019. *Europeana Pro* (blog). July 31, 2019. https://pro.europeana.eu/page/issue-13-ocr.

Fang, Zijin [方自金]. 2019.6. “Status Quo, Problems and Suggestions of the Photocopying and Publishing of Newspapers in the Republic of China” [民国报纸影印出版的现状、问题与建议]. *Information on Publication* [出版参考] 2019 (6). www.sohu.com/a/330753909_488898.

Fiormonte, Domenico. 2012. “Towards a Cultural Critique of the Digital Humanities.” *Historical Social Research / Historische Sozialforschung* 37 (3 [141]): 59–76.

———. 2016. “Toward a Cultural Critique of the Digital Humanities.” In *Debates in the Digital Humanities*, 2 (2016). Minneapolis, MN: Univ of Minnesota Press. https://doi.org/10.25354/nls2020.08.43.

“GRETIL - Göttingen Register of Electronic Texts in Indian Languages.” 2020. GRETIL - Göttingen Register of Electronic Texts in Indian Languages. June 26, 2020. https://doi.org/10.25354/nls2020.08.6.

Grüning, Tobias, Gundram Leifert, Tobias Strauß, Johannes Michael, and Roger Labahn. 2019. “A Two-Stage Method for Text Line Detection in Historical Documents.” *International Journal on Document Analysis and Recognition (IJDAR)* 22 (3): 285–302. https://doi.org/10.1007/s10032-019-00332-1.

Gülden, Svenja A., Celia Krause, and Ursula Verhoeven. 2020. “Digital Palaeography of Hieratic.” In *The Oxford Handbook of Egyptian Epigraphy and Paleography*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190604653.013.42.

Hockx, Michel, Joan Judge, and Barbara Mittler, eds. 2018. *Women and the Periodical Press in China’s Long Twentieth Century: A Space of Their Own?* Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108304085.

Li, Mingjie [李明杰], and Ruilong Li [李瑞龙]. 2016. “The General Trend, Practical Problems and Optimization Strategies of the Compilation and Publication of the Republic of China in the Past Ten Years (2005-2015)” [近十年(2005-2015)来民国文献编纂出版的总体趋势、现实问题及优化策略]. *Shanxi Archives* [山西档案], no. 05: 40–43.

Library of Congress and MODS Editorial Committee. 2018. “MODS User Guidelines (Version 3).” Metadata Object Description Schema - MODS. August 2, 2018. http://www.loc.gov/standards/mods/userguide/.

Magner, Thomas F. 1974. “The Latin Alphabet and the Languages of China.” *The Journal of General Education* 26 (3): 205–18. https://www.jstor.org/stable/27796437.

Mahony, Simon. 2018. “Cultural Diversity and the Digital Humanities.” *Fudan Journal of the Humanities and Social Sciences* 11 (3): 371–88. https://doi.org/10.1007/s40647-018-0216-0.

Mahony, Simon, and Jin Gao. 2019. “Linguistic and Cultural Hegemony in the Digital Humanities.” In *Proceedings of the Digital Humanities Congress 2018*. Sheffield, 6-8th September 2018. https://www.dhi.ac.uk/openbook/chapter/dhc2018-mahony.

Michaels, Axel, and Barbara Mittler. 2019. “Asia and Europe from a Transcultural Perspective - The New Heidelberg Centre for Asian and Transcultural Studies (CATS).” *International Institute for Asian Studies - The Newsletter* 82 (Spring): 44–45.

Mixter, Jeff. (2014) 2017. *VRA-RDF-Project*. XSLT. GitHub. https://github.com/mixterj/VRA-RDF-Project.

OpenDACHS - Centre for Asian and Transcultural Studies. 2020. “Open DACHS.” OpenDACHS - A Citation Repository for the Sustainable Archiving of Cited Online Sources. 2020. https://doi.org/10.25354/nls2020.08.35.

Philipps, A., and M. Davis. 2009. “BCP 47 - Tags for Identifying Languages.” Internet Engineering Task Force (IETF). https://doi.org/10.25354/nls2020.08.44.

Riley, Jenn. 2010. “Seeing Standards: A Visualization of the Metadata Universe.” Jennriley.Com. 2010. https://doi.org/10.25354/nls2020.08.45.

Smith, David A., and Ryan Cordell. 2018. “Report: A Research Agenda for Historical and Multilingual Optical Character Recognition (OCR).” Boston. https://ocr.northeastern.edu/report/.

Sturgeon, Donald. 2018. “Large-Scale Optical Character Recognition of Pre-Modern Chinese Texts.” *International Journal of Buddhist Thought and Culture* 28 (2): 11–44. https://doi.org/10.16893/IJBTC.2018.12.28.2.11.

———. 2020. “Digitizing Premodern Text with the Chinese Text Project.” *Journal of Chinese History* 4 (2): 486–98. https://doi.org/10.1017/jch.2020.19.

Sung, Doris, Liying Sun, and Matthias Arnold. 2014. “The Birth of a Database of Historical Periodicals: Chinese Women’s Magazines in the Late Qing and Early Republican Period.” *Tulsa Studies in Women’s Literature* 33 (2): 227–37. https://doi.org/10.1353/tsw.2014.0004.

The Unicode Consortium, ed. 2020. *The Unicode Standard, Version 13.0 – Core Specification*. Mountain View, CA: Unicode, Inc. https://www.unicode.org/versions/Unicode13.0.0/UnicodeStandard-13.0.pdf.

Visual Resources Association Foundation. 2006. “CCO Commons - Cataloging Cultural Objects (Pdf Version).” Guide. 2006. https://doi.org/10.25354/nls2020.08.46.

———. 2019a. “Cataloging Cultural Objects - About : What Is CCO?” CCO Commons. March 7, 2019. http://web.archive.org/web/20190307035239/http://cco.vrafoundation.org/index.php/aboutindex/.

———. 2019b. “Cataloging Cultural Objects - Examples.” CCO Commons. March 7, 2019. http://web.archive.org/web/20190307174324/http://cco.vrafoundation.org/index.php/toolkit/index_of_examples.

Volkmann, Armin. 2019. “Transforming Knowledge: Concepts of Transcultural Studies and Digital Humanities.” In *Engaging Transculturality*, edited by Laila Abu-Er-Rub, Christiane Brosius, Sebastian Meurer, Diamantis Panagiotopoulos, and Susan Richter, 413–26. Oxon: Routledge.

“VRA Core - a Data Standard for the Description of Works of Visual Culture.” 2018. Official Website. February 15, 2018. https://doi.org/10.25354/nls2020.08.3.

VRA Core Data Standards Committee. 2007. “VRA Core 4.0 Element Description.” Visual Resources Association. https://www.loc.gov/standards/vracore/VRA_Core4_Element_Description.pdf.

Wilkinson, Endymion. 2017. *Chinese History: A New Manual*. Fifth edition (Digital) Revised and Enlarged. Cambridge, Massachusetts: Endymion Wilkinson.

Xiao, Hong [肖红]. 2017. “The Main Problems and Solutions in the Practice of Digitization of Newspapers in the Republic of China” [民国报纸数字化实践中的主要问题及处理策略]. *Researches on Library Science* [图书馆学研究] 2017 (4): 22–37. https://doi.org/10.15941/j.cnki.issn1001-0424.2017.04.004.

Xiao, Hong [肖红], and Yan Huai [槐燕]. 2017. “An Analysis of Quality Checking Problems in the Practice of Digitization of Newspapers of Republic of China” [民国报纸数字化实践中的质检问题探析]. *Researches on Library Science* [图书馆学研究], no. 07: 61-78+87.

Xiao Hong [肖红], Wu Ming [吴茗], and Zeng Yan [曾燕]. 2015. “A Probe into the Digitization and Service of Microfilms of Newspapers in the Republic of China—Taking the National Library as an Example” [民国报纸缩微胶片数字化及服务探析——以国家图书馆为例]. *Journal of Library Science* [图书馆学刊] 37 (10): 89–92.

Zhang Wei [张玮]. 2019. “Research on the Common Problems in Digital Acceptance of Newspapers in the Period of Republic of China: A Case Study of National Library of China” [民国报纸数字化验收常见问题研究——以国家图书馆为例]. *Library and Information Studies* [图书情报研究], no. 3: 72–79.