International Journal of Digital Humanities (2023) 4:147–170
https://doi.org/10.1007/s42803-022-00055-6

ORIGINAL PAPER

Procedural challenges: the FAIR principles and PRC electronic resources - a case study of Chinese republican newspapers

Matthias Arnold[1] · **Duncan Paterson**[2] · **Jia Xie**[1]

Received: 17 June 2021 / Accepted: 14 November 2022 / Published online: 30 November 2022

Abstract

It is tempting to assume that FAIR data principles effectively apply globally. In practice, digital research platforms play a central role in ensuring the applicability of these principles to research exchange, where General Data Protection Regulation (EU) and Multi Level Protection Scheme 2.0 (PRC) provide the overarching legal frameworks. For this article, we conduct a systematic review of research into Chinese Republican newspapers as it appears in Chinese academic journal databases. We experimentally compare the results of repeated search runs using different interfaces and with different points of origin. We then analyze our results regarding the practical and technical accessibility conditions. Concluding with an analysis of conceptual mismatches surrounding the classification of items as "full-text", and of a case of total data loss that is nevertheless symptomatic of the limited degree of data re-usability. Our results show structural challenges preventing Findability, Accessibility, Interoperability, and Re-usability from being put into practice. Since these experiments draw upon our Digital Humanities (DH) research, we include a state-of-the-field overview of historical Periodicals and digitization research in the PRC. Our research on the one hand addresses DH practitioners interested in digital collections, and technical aspects of document processing with a focus on historical Chinese sources. On the other hand, our experience is helpful to researchers irrespective of the topic. Our article is accompanied by a data publication containing sources and results of our experiments, as well as an online bibliography of the research articles we collected.

Keywords FAIR data guidelines · OCR · Internet research · Chinese newspapers–History · Full-text databases · Digital resources of Republican China

✉ Matthias Arnold
arnold@hcts.uni-heidelberg.de

Extended author information available on the last page of the article

1 Introduction

Irrespective of funding outcome, during our research into Republican Chinese newspapers we frequently received *feedback* along the lines of: "I'm sure someone in China has done this already!" Reviewer#2 (Peterson, 2020). While statements of the form "someone in X has probably done Y" are invariably true, they remain unconstructive without the crucial details about who someone is, what it is that they have done, and how it is supposed to apply to the situation at hand. Because of its prevalence, we decided to systematically assess the assumptions that motivate Reviewer#2's opinion. At its core, these boil down to FAIR research practices being the global norm for conducting digital research (Hansen et al., 2018; Wilkinson et al., 2016). Research should be Findable (F1-F4), Accessible (A1-A2), Interoperable (I1-I3), and Reusable (R1-R3). We will take a detailed look at each of these criteria using our topic as a case study.

In our primary research area use of convolutional neural networks (CNN) for layout analysis and Optical Character Recognition (OCR) of historical Chinese documents is expanding the quality and scope of available sources (Liebl & Burghardt, 2020; Oliveira et al., 2018). It also places high demands on our ability to re-use electronic resources as training data. Research in this area is fast-paced, and heavily relies on dissertation work. Lack of accessibility is therefore painful, potentially leading to duplication of work. Yet, without means of reusing previous work, such duplication remains a necessity. We can therefore compare research in this area between historical newspapers in Latin and CJK (Chinese, Japanese, Korean) (The Unicode Consortium, 2020) scripts. While the problems, such as complex layouts, rare glyphs, or volatile primary sources are largely identical, the means to build upon previous research is not. Contrasting the state-of-the-field in both linguistic areas serves to show how procedural challenges and conceptual difficulties can prevent data re-use between the EU and the PRC. This section also serves to present a summary of our findings to researchers working on other complex CJK documents. We supplement this written account, which is invariably going to be outdated between submission and publication, by both a data repository and an online bibliography we maintain.

When looking at OCR as a key technology (Smith & Cordell, 2018) for Digital Humanities (DH) we see both a presentist and a Latin bias within specialist and general tools. As automating the conversion from binary (images) to text formats remains crucial to digital collections at both large (Sturgeon, 2017) and small scales (Tu et al., 2020) it deserves special attention. Lastly, the availability of high-quality full-texts is critical to the creation of linguistic corpora (Yasuoka, 2019; Zhang & Xue, 2012) for Natural Language Processing (NLP). In the field of OCR, we encounter another popular myth about the over-complexity of CJK scripts. Though the over 90.000 CJK Unified Ideographs defined in Unicode dwarf for example Latin, which has roughly 1300 with all extensions and 52 without, or any other script, it is sure Latin bias exists when popular tools do not support CJK by default, when CJK textual phenomena do not appear in global encoding standards and so on. As our review of OCR research in China shows,

the overcomplexity of the Chinese script is not the primary driver for this. CJK OCR tasks are often computationally more expensive since they have to cover a larger result space than Latin scripts, but the power of contemporary computer tools is more than adequate to perform such tasks in either case. To understand the sources of Latin bias, we must look elsewhere.

Our article addresses two audiences simultaneously. On the one hand, we wish to provide a structural overview of the quickly evolving state-of-the-field. As our literature review below shows, achieving a comprehensive perspective on research conducted in greater China on any given topic is systematically hindered by several bottlenecks which apply irrespective of the research topic. On the other hand, we hope to alleviate such problems for those working on similar materials, so they can refine their search methods, and find noteworthy research directly from here or the accompanying materials that we provide.

1.1 Method

For this article, we conduct a literature review of research into Republican Chinese newspapers as it appears in popular Chinese academic journal databases and their accompanying data publications. In essence, we repeatedly run identical keyword searches using different interfaces and compare results. We first gather materials in the form of saved HTML files and screenshots which are then prepared for further analysis. The source files are accessible on Zenodo at https://doi.org/10.5281/zenodo.6801936, along with descriptive metadata and additional sources not used in the final article version. All data is collected on Windows 10 Enterprise Edition, using Firefox Browser 88.0.1, and cross-checked with both Google Chrome 90.0.4430.212 and Microsoft Edge 90.0.818.62.

Our paper targets several academic platforms accessible via *CrossAsia*, a subscription service hosted by *Staatsbibliothek zu Berlin*.[1] It contains 133 databases in and about East Asia in CJK languages. For comparison with *CrossAsia*, we use two Chinese VPNs, one academic associated with Zhejiang University (浙江大学) and one commercial purchased from the Chinese online shopping platform taobao.com. Both VPNs consistently returned identical results for academic platforms. To ensure comparability we restrict our searches to platforms accessible both via VPN and *CrossAsia*. We, therefore, had to exclude frequently-used platforms Chaoxing Qikan Library (超星期刊) and Weipu Wang (维普网) since *CrossAsia* does not subscribe to them. We also ruled out National Social Science Database (国家哲学社会科学学术期刊数据库) of Chinese Academy of Social Sciences, which is open access but possesses far fewer resources than the ones below. The resulting platforms are: *China National Knowledge Infrastructure* (中国知网) CNKI, *Wanfang Data* (万方

[1] CrossAsia, "About CrossAsia," 2021, https://blog.crossasia.org/about/?lang=en is one of the portals of the Specialized Information Service Asia (FID Asia), funded by the German Research Foundation (DFG); cf. Marietta Fuhrmann-Koch, "DFG fördert neue Fachinformationsdienste für Asien und die Altertumswissenschaften," news, idw - Informationsdienst Wissenschaft, February 16, 2016, https://idw-online.de/en/news646187.

数据), and *Airiti Library* (華藝線上圖書館). CNKI, run by Tsinghua University and Tsinghua Tongfang Company (清华同方), is the largest aggregator and distributor of digital resources in China that claims to have integrated over "95% of all Chinese academic resources." CNKI has built its integrated knowledge resources database since 1996.[2] Compared to CNKI, *Wanfang Data*, an affiliate of the Ministry of Science & Technology of China (中华人民共和国科学技术部), includes fewer types of journals with a smaller time frame and a slightly slower update frequency. Besides, *Airiti Library* mainly provides academic papers in Taiwan. Furthermore, Quanguo baokan suoyin 全国报刊索引 (CNBKSY) by Shanghai Library (上海图书馆), also known as Shanghai Institute of Science and Technology Information (上海科学技术情报研究所), is the leading provider of Republican period materials.

VPNs let us compare different versions for national (https://www.cnki.net/) and overseas (https://oversea.cnki.net/) audiences for CNKI. Each of which provides a configurable GUI option for either "English" or "Chinese" menus. Users selecting to view CNKI via the English GUI are automatically rerouted to the overseas version. The reverse however is not true, an alternative Chinese interface of the overseas version (https://chn.oversea.cnki.net/) is offered instead. In addition, platforms like CNKI or CNBKSY feature complex package models for licensing their sub-databases. In practice, searchable databases, user interface, and default download formats differ between these versions.

To contextualize our initial results, we review the academic literature with an eye toward the tools they mention, and towards reproducing their results from overseas. This is based on a selection of articles that we share as Zotero group library "Procedural_challenges" at https://www.zotero.org/groups/4046522. Lastly, we reach out to individual researchers, and practitioners involved in e.g. *CrossAsia*, as well as two social media groups: The invitation-only DH WeChat group "数字人文2群 | DH Group2" with 369 members mainly from the PRC, and the private "Digital Sinology" Facebook group with 1999 members from around the globe. The discussions and polling we conducted in these groups greatly assisted our research and are part of the supplemental data repository.

2 Findable

To locate research from the PRC in general, and about digital methods relevant to historical newspapers in particular, we have to look at globally unique persistent identifiers (FORCE11, 2014). Such IDs are catalogued and made searchable by research data platforms. We start by focusing on the practical demands that these platforms accurately return all such IDs to allow us to learn about the existence of relevant research materials in the first place. We will ignore the technical evaluation of ID schemas or standards within the results.

[2] "收录了95%以上正式出版的中文学术资源" 'CNKI Introduction', accessed 11 January 2021, https://enscholar.cnki.net/home/about.

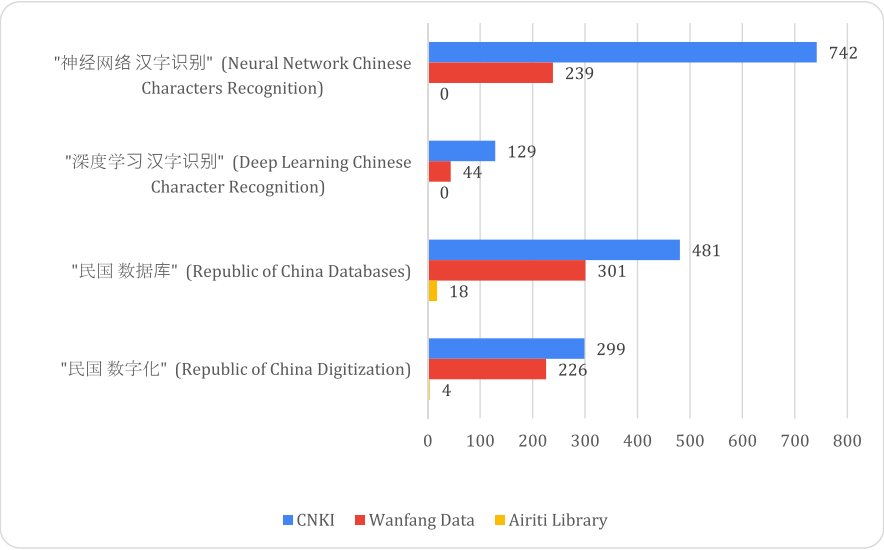

Fig. 1 Comparison of search results in CNKI, Wanfang Data, and Airiti Library

2.1 Chinese academic platforms

We perform identical searches in CNKI, *Wanfang Data*, and *Airiti Library* via *CrossAsia* for four keyword groups related to our research, within "All" databases of each platform (Fig. 1):

- "神经网络 汉字识别" (Neural Network Chinese Character Recognition),
- "深度学习 汉字识别" (Deep Learning Chinese Character Recognition),
- "民国 数据库" (Republican China Databases), and
- "民国 数字化" (Republican China Digitization).

CNKI consistently returns the most unique search results, while *Airiti Library* contains the fewest relevant items. In our search, CNKI covers more journals, such as Library Tribune, which are not included in *Wanfang Data*. In addition, more results appear in CNKI for journals that are part of multiple platforms, such as Researches on Library Science. At this point, CNKI is the best service for our case study, as it indexes the most identifiable objects. Given the dominance of CNKI search results, we continue to focus on it as the main source of literature in this paper. Because CNKI has different versions, we need to compare the results presented by each version, including reproducing searches mentioned in relevant articles, to better understand the consistency of CNKI search results.

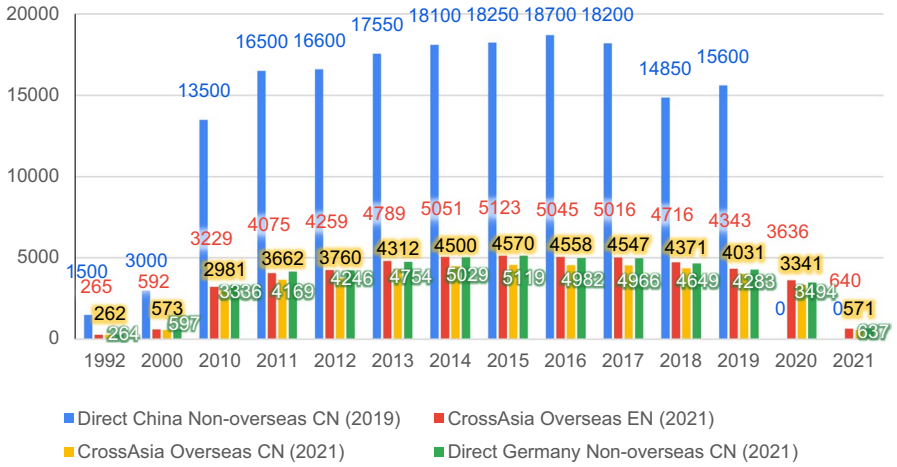

Fig. 2 Subject (主题) search: "民国" (Republican China) in "总库" (All databases)

2.1.1 Different search results in CNKI

When we try to reproduce the numbers presented by (Li & Fang, 2019, p. 92 fig. 1) through *CrossAsia* we receive drastically different results. Their discussion of a CNKI-generated graph about papers related to "民国" (Republican China) features more than 5.000 publications in 2003 rising to 17.500 related articles each year between 2013 and 2017. However, configuring to retrieve all available publications (databases "All"), our results are markedly different.

Figure 2 compares the approximate results from Li and Fang (2019 fig. 1) in blue (January 2019), with those conducted by us via *CrossAsia* of the overseas CNKI, English in red, Chinese in yellow, and the non-overseas version with direct access from Germany in green (all May 20, 2021). Related publications only exceed the 5.000 mark between 2014 and 2017, but remain significantly smaller throughout. While we would expect some variation of results based on new articles being continuously published, embargoes expiring, and with occasional removal of erroneous data, none of these processes can explain the observed differences in total hits. Thus, we began experimenting with different interfaces and means of access. Changing the interface of overseas CNKI from English to Chinese, we get different numbers. Surprisingly, the results with the Chinese interface are slightly smaller than with the English one. When repeating our search in the Chinese interface on the non-overseas server without VPN, the numbers are in between either language setting for the overseas search. While we do not expect results to be static, the variation within a short time frame is surprising. Especially the variance of up to 10% based on GUI language preference is unexpected.

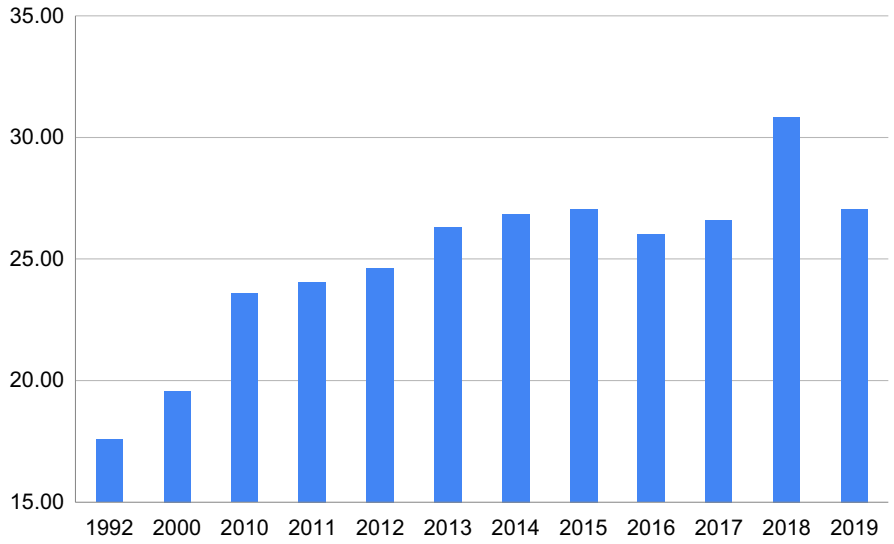

Fig. 3 Average of results compared to (Li & Fang, 2019) in percent

Nevertheless, these variations are negligible compared to the difference between our search results and those by Li and Fang.[3] Figure 3 shows that there is not a single year where we receive as much as a third of their results, and less than 20% for older records.

To verify the dimension of the problem, we asked social media users to repeat our search and share their numbers, allowing us to gather data from different access locations. One respondent, Ms. Tang Li, a Chinese Studies Librarian at the University of Southern California points to the list of "restricted titles" within CNKI, that are inaccessible to overseas users. The list from December 2018 comprises 10,474 journals, of which 150 are "restricted".[4] Professor Adam Smith of the University of Pennsylvania provided us with his numbers which peak in 2015 but remain below the 50% mark of our results.

What does that mean? The variation between our test runs could point to lingering network effects, where significant portions of the collected data become temporarily unreachable and are no longer returned by the search interface. The IDs might be eternal (F1), but that is insufficient to ensure findability when the registering of the ID within the index (F3) is not. CNKI statistics appear unstable and should be contextualized by license, version, GUI options, etc. for reproducibility. It

[3] The authors mention that they collected the data in January 2019, therefore the numbers they give for that year are a prediction.

[4] The list (EastView, 2018) is linked from Eastview's China Academic Journals (CAJ) fact sheet, https://www.eastview.com/resources/journals/caj/. However, the journals listed there are probably restricted due to licensing agreements (English language editions of Chinese journals), their military nature, or are about geological resources. Therefore, these restrictions do not affect the results in question here.

is both counterintuitive and undocumented, that changing the GUI language preference results in ±10% search hit variance. Systematic filtering is likely responsible for such effects. But we are unable to determine who is responsible for such filtering and for what purpose. Politically motivated filters for Chinese webpages are a known aspect of the "Great Firewall" (Wu & Lam, 2017). As are deletions of articles based on shifting political criteria (Tiffert, 2019). Either one could in theory account for variation within our results. How or why large swaths of Republican period secondary scholarship would suddenly trigger those mechanisms depending on GUI language and search time remains hard to imagine. Yet, neither effect can explain the massive differences between the published figures, and our results. It is theoretically possible that changes to the indexing configuration between the time of Li and Fang's search and our own have drastically altered the total number of hits. Similarly, large-scale deduplication could partially explain what we are seeing. In either case, we would expect some information about such changes to be provided either by CNKI or by responsible individuals. In fact, in 2020 (precise date unknown) CNKI implemented a series of changes to its technical infrastructure, including both search engine and user interface upgrades. Could this explain the variance? The official blog about the KNS (Knowledge Network System)[5] update, however, shows that users should only expect more results, in their case an increase from 16,687 to 21,128.[6] This increase is in addition to a previous update to KNS subject search in 2018 fixing incomplete results.[7] It seems unlikely that our particular search reverses the documented outcomes of these backend updates, similarly updated GUI components should not impact the total number of search results, however they are displayed.

We have seen the need to compare multiple runs of the same search within CNKI, which leads to questions about the accessibility of CNKI as a service, and the items it presents. Comparing the results for both the overseas version accessed through *CrossAsia* and the mainland version accessed through VPNs. For example, the search for "Republican Digitization" returns 340 journal articles and 210 theses, irrespective of academic or commercial VPN, via *CrossAsia* we receive 333 and 213 respectively. Visiting the mainland version of CNKI through either VPN returns consistent results, yet results differ for the overseas version depending on language setting and are never identical to the mainland version.

[5] KNS (Knowledge Network System) is the basis of CNKI platform, see the user guide to an older version: https://staatsbibliothek-berlin.de/fileadmin/user_upload/zentrale_Seiten/ostasienabteilung/pdf/UserGuide35.pdf.

[6] "CNKI Has Updated? How to Search in New CNKI?" (知网 (CNKI) 更新了？新版知网如何进行文献检索?), accessed February 10, 2022 https://www.sohu.com/a/403785796_100020119.

[7] "Announcement on the Comprehensive Upgrade of CNKI Search System" (关于中国知网智能检索系统全面升级的公告), accessed February 10, 2022 https://bianke.cnki.net/pulpit/Details/index/2298.

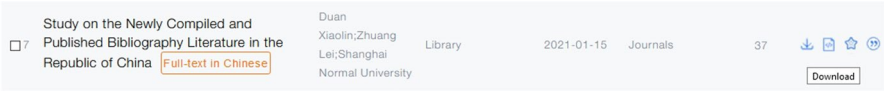

Fig. 4 Example of "download"

3 Accessible

While the FAIR principles focus on technical aspects of ensuring accessibility through communication protocols, we continue to focus on questions of practical accessibility for CNKI resources. As shown, findability within CNKI, irrespective of technical protocol, needs to be taken with a grain of salt. Then what about the items we do find in different versions, what goes into determining their accessibility, essentially our ability to view or download not just the catalogued metadata but the actual item itself. For the same items, how does accessibility compare between versions of CNKI? We try the same search as Li and Fang (2019) via academic VPN to match their original setup. Very few items are unavailable for download, so we assume the same holds for their search. To further test this, we expand our original list of keywords using CNKI recommended subject headings. We add three general search terms:

- "民国报纸" (Republican newspapers),
- "民国资料" (Republican materials), and
- "民国文献" (Republican literature).

For the Chinese character recognition aspect, we run combined searches for "汉字识别" (Chinese character recognition) together with:

- "神经网络" (neural networks),
- "卷积神经网络" (convolutional neural networks),
- "深度学习" (deep learning), and
- "机器学习" (machine learning).

Lastly, we add two broader search terms related to DH:

- "数据库" (database), and
- "数字化" (digitization).

Combined with the recommended headings "民国", "民国报纸", and "民国资料", we run 14 subject searches within "Academic Journals" (Fig. 5) and "Theses & Dissertations" (Fig. 6). To assess the accessibility, we then compare the number of active download links to the total number of results.

Downloadable items show the "Download" icon (Fig. 4). Counting those determines the actual number of downloadable items.

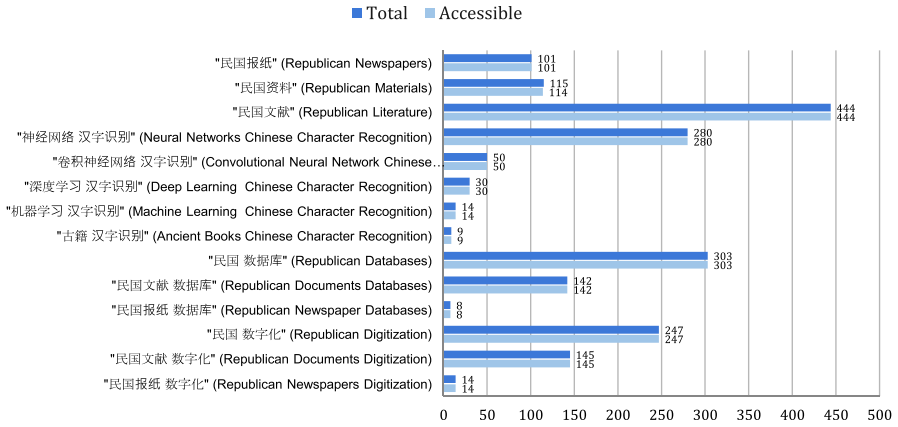

Fig. 5 No gap between accessible and total search results within "academic journals"

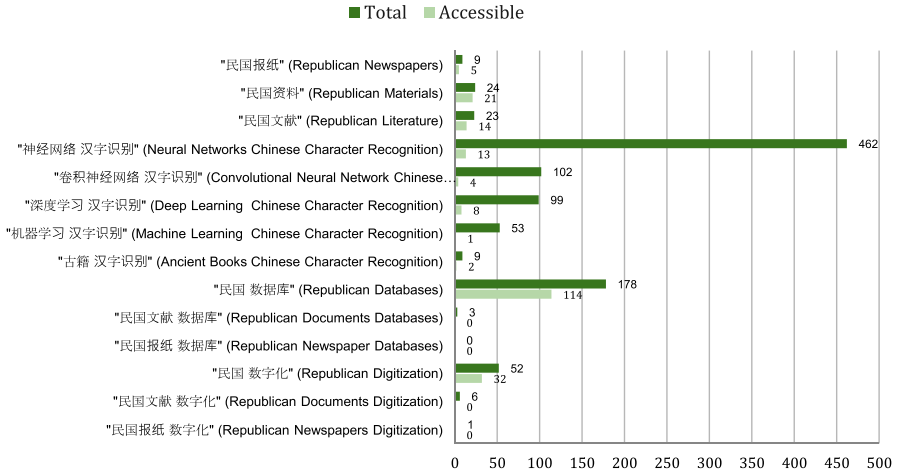

Fig. 6 Big disparity between total and accessible search results within "theses & dissertations"

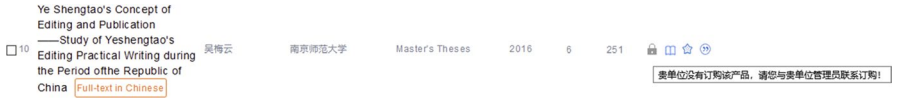

Fig. 7 Example of "no subscription"

Fig. 8 Example of "prohibit"

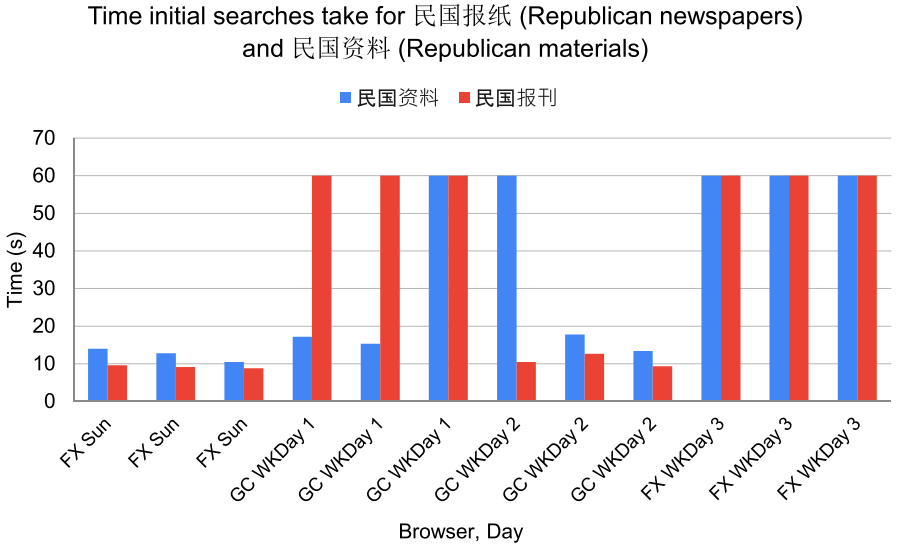

Fig. 9 Time for initial searches

In the mainland version, all articles and theses are also downloadable. In the overseas version via *CrossAsia*, all but one journal article are accessible (Fig. 5). Yet, within "Theses and Dissertations" out of 1021 hits, only 20.96% (214) are available (Fig. 6). Unavailable items come in two flavors with distinct error messages. "Your organization has not ordered this product, please contact your administrator to order it. (贵单位没有订购该产品, 请您与贵单位管理员联系订购)" (Fig. 7), pointing to licensing issues. In fact, the *CrossAsia* subscription of the "Dissertations" database only covers 4 out of the 10 available series. As a workaround, they suggest 中国学位论文全文数据库 (China dissertations database) by *Wanfang*. We once observed "Prohibit" which seems to indicate a more general restriction (Fig. 8).

Since all results are available from within China, we can preclude embargoes skewing results. Regrettably, alternative means of access via *CrossAsia* are not immediately visible to the end-user, but fortunately, they exist and we seem to get all titles irrespective of search route.

Lack of access is particularly painful for our work on computational methods where dissertation-level work frequently forms the bases of subsequent commercial and proprietary follow-up implementation. Therefore, even if a dissertation relevant to our research exists, without being able to access and evaluate it, such knowledge is of limited practical use. The problems of practical accessibility, however, do not stop at questions of licensing and embargos, they also include technical aspects of network connectivity.

3.1 Technical problems

We frequently encounter network connectivity issues during our experiments. Unstable network environments and slow loading speeds are a particular problem

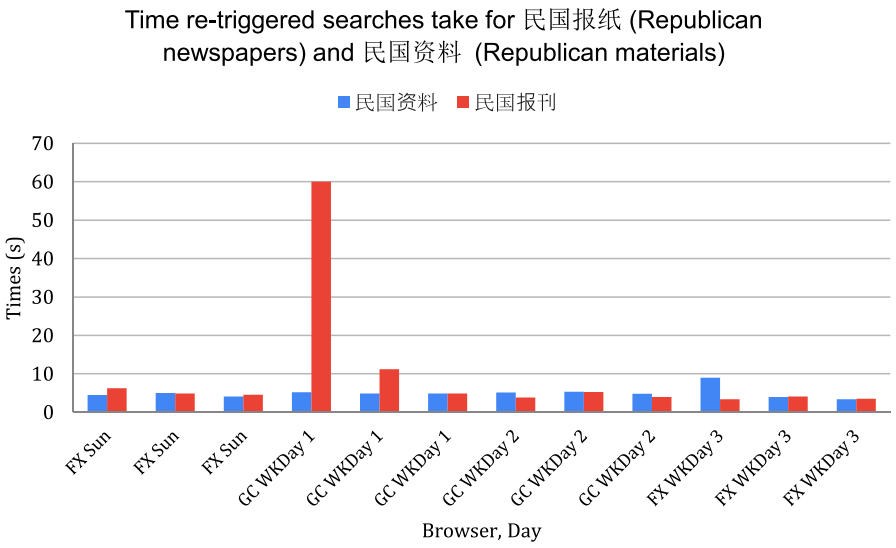

Fig. 10 Time for re-triggered searches

for access via *CrossAsia*.[8] We tested at different times of the week (Sunday, three workdays) and times of the day to control for overall network traffic, with Firefox (FX) and Google Chrome (GC) to control for browser issues. We measure the initial time between submitting searches for "民国资料" and "民国报刊" and loading the results, as well as for retriggering searches. Wait times over 60s constitute a timeout (Figs. 9 and 10).

Two patterns become visible: Firstly, initial searches from CNKI's homepage during CET working hours are indeed slow, and frequently time out. Subscribers can expect a better quality of service on Sundays. Secondly, once a search is successful, some form of caching seems to ensure that searches triggered from the result page of a previous search no longer time out. Users should therefore run as many searches per session as possible. This contrasts sharply with the experience of running searches via the academic VPN to CNKI's main server, where we see overall better performance. For a detailed analysis, we pinged the overseas server directly since doing so via the CrossAsia version consistently timed out. The average time for a direct connection is 244.8ms compared to 1.4ms via VPN, for comparison the connection to CrossAsia's portal page is 21.8ms.

The tests show that these effects are not related to the infrastructure of CNKI itself. Our questions to *CrossAsia* about their network infrastructure and traffic monitoring, unfortunately, remain unanswered. We can show that while the servers do not suffer from connectivity issues or high load when accessed via alternative routes, access via *CrossAsia* frequently times out, and displays high load times,

[8] as visible from the large number of complaints on their forum https://forum.crossasia.org/c/crossasia-datenbanken.

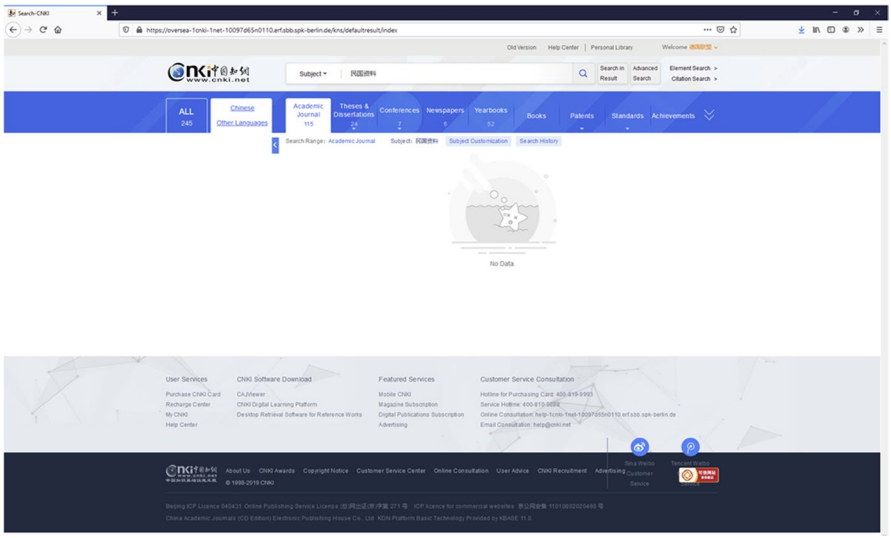

Fig. 11 Screenshot of "no data"

leading to empty search results (Fig. 11). Based on our tests, complaints in the *CrossAsia* forum are well founded.

Figure 11 shows there to be "no data" for "民国报纸", which is false. Connectivity issues are obscured from the user. Thus, poor GUI design from CNKI might lead researchers to bad conclusions, namely that no results match the given search term. This is confounded by the fact that *CrossAsia* does not provide an endpoint to monitor connectivity. In cases where network connectivity is to blame for lack of results, this should be indicated to the user, so they know to try again at a different time.

3.2 Open access and digital resources

So far, we looked at the operations of large research platforms. This ignores the accessibility of individual DH projects. With academic literature, the core accessibility question is if the contents in question are retrievable or not. For other forms of research output, such as databases of Republican resources in China, accessibility is a more complex topic, with endless potential restrictions to contend with. We select three articles that focus on databases of Republican materials in China and that provide a state-of-the-field type summary:

- (Duan, 2016) introduces the digitization of Republican materials since 2010, listing databases with their features and shortcomings.
- (Wu, 2017) summarizes DH projects in the digitization of Republican literature, including databases, tools, and methods.
- (Luo, 2020) discusses the Tianjin Library database as a representative case of historical newspapers.

Table 1 Operator and count of databases

Operator	Count
Academia Sinica	2
PRC university Co-op	2
PRC university libraries	8
Missing	8
Special thematic focus	9
PRC public libraries	20
Sum	49

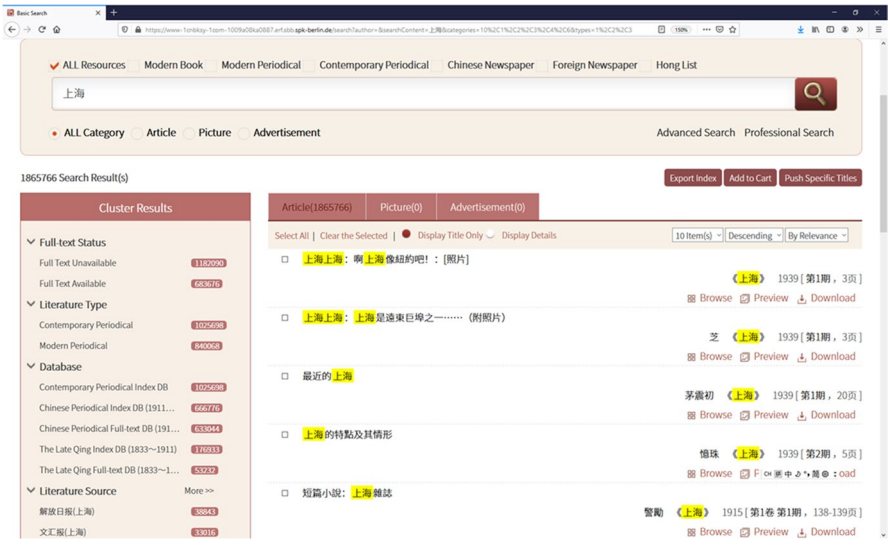

Fig. 12 Search results for "上海" (Shanghai) in CNBKSY via CrossAsia

We attempt to access these databases one by one (Table 1), to understand the availability from Europe, while compiling a catalog of all mentioned resources including their URL, type, accessibility, availability of full-text, etc.

The 49 databases cover periodicals, images, and videos. However, only two are open access: the image database of modern literature (中国近代文献图像数据库) and the database of Northeast Anti-Japanese United Army (东北抗联数据库). Ignoring the 8 missing resources, all others implement access restrictions.

When Chinese libraries require a user account for access to free resources, users need a Chinese mobile number, effectively excluding international users. For example, the Republican China Literature (民国时期文献) explicitly limits the registration to Chinese citizens. Technical limitations exist as well. Network issues consistently limit access to the deep text mining system All Tang Poetry Analysis System (全唐诗分析系统), while some pages of Beijing Normal University Library require the defunct Flash package, making them inaccessible to modern browsers. The

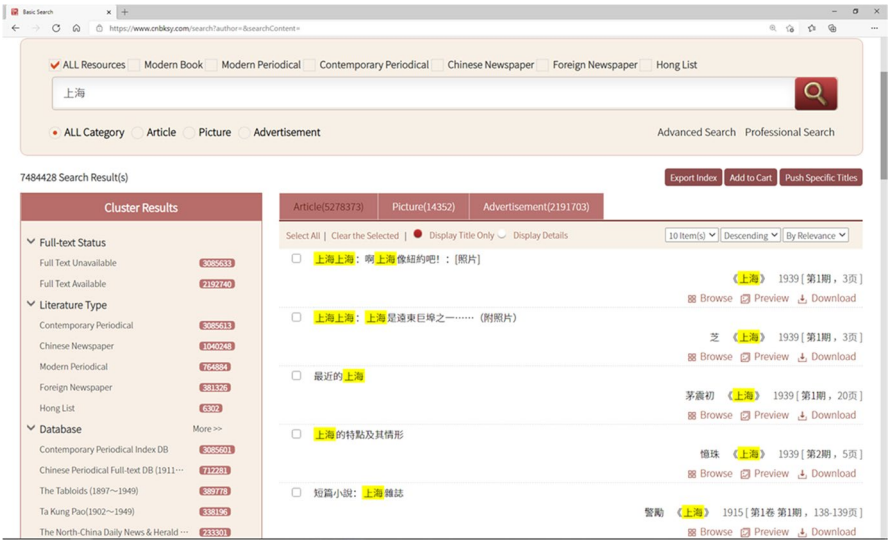

Fig. 13 Search results for "上海" (Shanghai) in CNBKSY via VPN

remaining databases require subscriptions. Those run by libraries usually restrict access to local users. With *CrossAsia* we can access three more: China Academic Digital Associative Library, Dacheng Old Periodicals database (大成老旧全文数据库), and CNBKSY. However, *CrossAsia* only includes three licenses of the over 20 databases and sub-databases with CNBKSY.

Taking CNBKSY as an example, we again compare access through *CrossAsia* and academic VPN. This should enable us to see if location-based filtering occurs on the level of individual research databases as well. A search for "上海" (Shanghai) returns 1.87 million search results via *CrossAsia*, roughly 25% of the 7.48 million via VPN.

We compare the unsubscribed sub-databases by looking at the side-bar in Figs. 12 and 13. The *CrossAsia* search retrieves five database collections, namely the *Contemporary Periodical Index DB*, and two each of the late Qing and Republican periodical databases (index and full-text), without others like The Tabloids (1897–1949). Nothing is on display in the "Picture" and "Advertisement" categories.

CNBKSY is a prominent example of how data providers monetize resources. Although the platforms provide a homogenous interface, the contents require individual licensing. The exact configuration of an institution's license package is hidden from end users. This kind of outsourced commercialization of publicly held data is common among Chinese historical resource databases preventing open access to public data. Examining the accessibility from outside the PRC, even fewer resources are openly accessible, due to further technical problems, licensing issues, and indirect requirements. Such limitations raise problems for computational research where scale is essential. *CrossAsia* has made some initiatives beyond the traditional subscription-based model to provide readers with more access at scale, such as *CrossAsia* Integrated Text Repository (ITR) and *CrossAsia* full-text search. To enhance

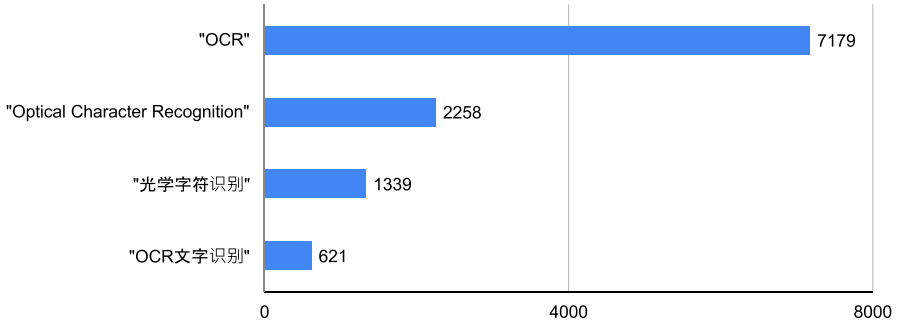

Fig. 14 Search results of "OCR" and its Chinese translation

accessibility, libraries can move beyond current subscription models by offering open access data to the public.

4 Interoperable

To be interoperable we should "use a formal, accessible, shared, and broadly applicable language for knowledge representation" (FORCE11, 2014). Noticeably this principle (I1) does not specify machine or human language. For our current purpose, we focus on the role of human language search terms, as they find their way into (semi-) automatically generated vocabularies. Technical terms such as "OCR", or "database", are used interchangeably with their Chinese translations within the literature under review. Since interoperable keywords use controlled vocabularies, it is technically trivial to match English language articles with the subject heading "China" to Chinese articles with the subject heading "中国" (*Zhongguo*). We will look at two scenarios where linguistic preferences impact the research process. First, we compare the results between searches for English vs. translated terms. Afterward, we take a closer look at how the term "全文" (*full-text*) is applied differently depending on natural language context. In both instances, awareness of natural language conventions plays an important, and so far underappreciated, role in successfully conducting online research. While Sinologists are often well aware of differences between 全文 resources, specialists for machine learning may have a very different set of expectations as to both the meaning of the term and what is implied by full-text resource collections when looking for training data.

4.1 English terms vs. Chinese translations

Besides author-supplied keywords, CNKI categorizes submissions with its subject headings. These seem automatically generated and show a mix of Chinese and English words, as well as acronyms for English terms. Since there is no reason why popular keywords in one language should not automatically be matched to another, we ask if we get identical hit counts irrespective of keyword language. Or, is it

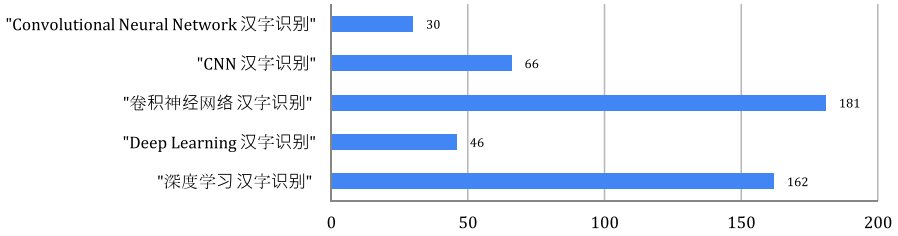

Fig. 15 Search results of "CNN" and "deep learning" and their Chinese translations combined with "Chinese characters recognition"

transparent to users searching for Chinese translations that terms in English yield better results, and is there a transparent method for which articles are categorized with Chinese vs. English terms? To find out we compare searches for technical terms in English with their Chinese translations.

"OCR" yields 7179 results, as opposed to 1339 for "光学字符识别" (Fig. 14). However, hit counts for "CNN" (Convolutional Neural Networks) are less than for the Chinese term "卷积神经网络". The same applies to "Deep Learning" and "深度学习" (Fig. 15). The first challenge when using a huge digital platform like CNKI is to find the terminology for optimal search results. With the mix of English expressions and Chinese translations, this is far from trivial. Users must effectively know the best search terminology in advance. Even within Chinese, as Prof. Smith points out, there are different results when searching with *fantizi* 民國 (1633 records) instead of *jiantizi* 民国 (2428), or Pinyin "*minguo*" (20). This implies that no automatic character conversion is applied here, where characters will be converted to e.g. simplified characters, whichever version the user enters. Similarly, the controlled vocabularies for subject headings do not make linguistic cross-references. From a technical standpoint these are solved problems, here would be a good opportunity to assist both Chinese native and non-native speakers alike, by transparently using GUI language preferences to configure search matching.

4.2 Full-text vs. 全文

For book-type resources, full-text databases exist since the early 1990s thanks to the ongoing digitization efforts of large PRC libraries (Tsui & Wang, 2020). For newspapers and other periodicals, the conversion started in earnest in late 2008 (Wang, 2014; Xia & Bao, 2020). As we shall see, the start time of a digitization project is important for the question of what full-text entails. Taking two examples from CNBKSY, 民国时期期刊全文数据库(1911~1949) (Chinese Periodical Full-text Database) and 晚清期刊全文数据库(1833~1911) (Full-text database of late Qing periodicals), both include "full-text" (全文) in their name (Guo & Ren, 2018). Upon closer inspection, these databases offer an author-title index as well as a basic classification of items, but no full-text. Connecting image scans to a database of titles is certainly useful, but also very clearly not a full-text transcription of the whole scan or article.

Fig. 16 Screenshot from CNBKSY showing full-text database records

Based on our review of the available literature about individual database projects, and especially by reaching out to practitioners via social media, we discern two sources for the terminological mismatch. The first concerns the way that CNKI and CNBKSY label databases. In CNKI one can assume that most contents labeled as full-text is, while for CNBKSY it is not. The latter does construct full-text databases, but these are labeled either "(OCR)" or "(Full-text)" (Fig. 16). With the versions accessible to us, we are unable to verify their content. But from a screenshot received through private communication in the WeChat group, we know they exist.

Whether a database contains full-text or not is largely left to the user to find out. Databases that contain full-text may be advertised as such, although access to full-text searches is unavailable. It is also possible that what counts as "full-text" has significantly changed during the last 20 years, which is the second source of confusion.

Similar to the West, it was a milestone in database development to provide digital facsimiles in addition to metadata referring to reprints or microform versions of resources. Soon, many systems providing more than image scans, like author/title indexes, were labeled "full-text". This also applies to China, where "全文" is the correct literal translation of "full text". However, while the use of "full text" nowadays describes the existence of a machine-readable text, this is different in the Chinese context. In some cases, image scan databases were re-labeled to "全文影像数据库" (full-text image database), or in the case of CNBKSY are called "全文数据库" (full-text database) because they plan to provide both, full-text and image scans. The fact that all of these databases are labeled "全文" reinforces our earlier point about the benefits of multi-lingual controlled vocabularies and shared terminologies. Otherwise, Reviewer#2 might have good reason to believe that a full-text version of a resource already exists, when in fact it refers to image data taken from microfilm.

5 Re-usable

Re-usability ties the previous principles together: findability, interoperability, and accessibility enable re-usability. It is also technically and practically the most difficult principle to implement as it touches upon various aspects beyond the direct control of individual researchers or collaborative projects, like licensing and community standards. Therefore, we continue our focus on what practical hurdles we

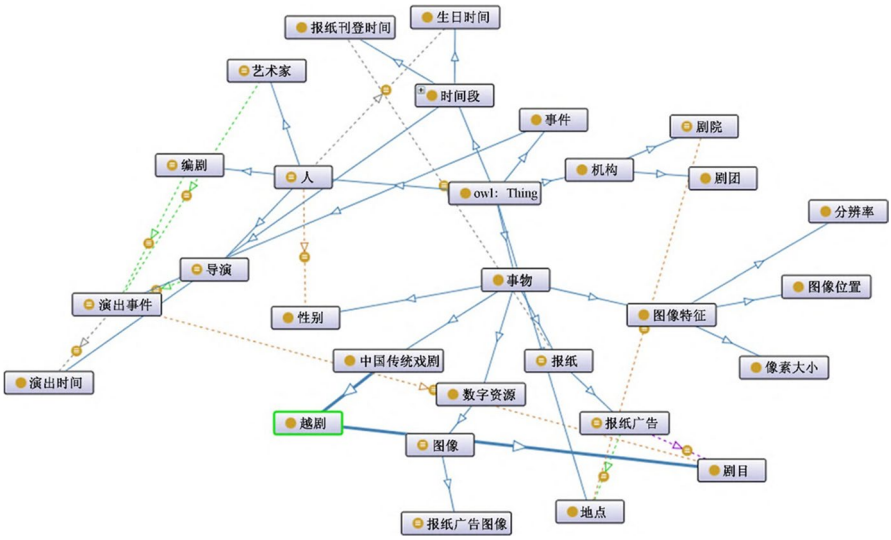

Fig. 17 Figure 6 in Yang and Xu, 2020 showing a visualization of the "Ontology for newspaper advertisements to Shaoxing Opera"

encounter when trying to re-use research data from the PRC, looking at one example that promises to embody best practices.

The article *Semantic Labeling of Advertising Images in Republican Chinese Newspapers* (Yang & Xu, 2020) stands out for being a study primarily using data extracted from a Republican newspaper, and for its state-of-the-art methodology. Our experience with its data serves as an example of all-too-common issues. The authors analyze *Yueju* 越剧 (Shaoxing Opera) advertisements from the Xinwen Bao 新闻报 (1923–1949). *Yueju* underwent a transition from its first introduction in 1917 to the city's most popular opera genre in the late 1930s. Analyzing changes in 2818 advertisements the authors identify four periods of *Yueju* development. The data set underlying their analysis was published and registered with a handle.net ID. In the paper, they stress the importance of semantic modeling and linked data in the DH, as a basis for their ontology models. In other words, their work relies on the kind of re-usability that interests us, making this article a good fit for our needs.

Based on the data extracted from CNBKSY they discuss the depth of metadata provided in that database and argue that especially the subject categories are not sufficient to analyze opera advertisements. Therefore, they develop their own "Ontology for newspaper advertisements to Shaoxing Opera" shown in Fig. 17.

To annotate the advertisements, they set up an IIIF service and use Mirador to add bounding boxes and assign labels from their ontology (Fig. 18). They are thus able to directly add in-depth annotations related to their research questions. Based on these annotations they define several phases in the development of *Yueju* advertisements and visualize the results, for example with a chart of the numbers of *Yueju* advertisements distributed over time (1926–1949).

Fig. 18 Figure 9 in Yang and Xu, 2020, showing the annotations interface in Mirador

This research-driven approach, the compilation of a specific data set, and the development of a domain-specific ontology to analyze the data represent best practices using appropriate standards and widely accepted tools. The published version of the article was submitted in October 2019 and revised in January 2020, it appeared on June 30, 2020, in the "online first" section of the *Library Journal*.

In the fall of the same year, the handle link did not work anymore: "The handle you requested cannot be found". We could, however, still contact the authors directly via email. They informed us that they conducted their research on a personal laptop and the device suffered an unrecoverable hard disk error. Why the registered handle link does not work remains unknown, but the result is that all the data is irretrievably lost: image scans, annotation data, and the ontology itself.

While complete data loss is the exception, this case is indicative of a wider problem. In our experience dead or missing links are unfortunately the norm, and few publications make the effort to publish their data in the way the authors attempted to. The fact that despite such favorable circumstances we still cannot use this data points to a deeper problem. In the humanities, data re-usability remains a rare exception compared to other disciplines. While national infrastructure changes, like the Dataverses 北京大学开放研究数据平台 (Peking University Open Research Data Platform), launched 2015, and the DataSpace@HKUST at Hongkong University of Science and Technology launched 2017 are beginning to take hold and can address researchers concerns with the Multi-Level Protection Scheme 2.0 (MLPS2.0) compliance, there is still much room for wider adoption. Infrastructure changes alone without accompanying changes in best practices and research procedures will be insufficient to combat the re-usability crisis. Infrastructure changes alone without accompanying changes in best practices and research procedures will be insufficient to combat the re-usability crisis.

6 Conclusion

Research platforms are in a crucial position to enable or hinder academic best practices. It is therefore necessary to critically evaluate their performance towards that goal. Simply assuming that FAIR principles apply does not take the practical, legal, and technical challenges for global research into account. Our experiments show the multi-faceted difficulties of international research data exchanges between the PRC and the EU. Framing these problems within the FAIR principles demonstrates the practical challenges preventing findability, accessibility, interoperability, and subsequently re-usability. Platforms such as CNKI provide useful functions in the form of full-text search and analytical tools for the quantitative handling of search results. This level of transparency, however, is built on a foundation of inconsistent results, and even intentional obfuscation (Tiffert, 2019). As such, the implications of our findings go beyond the relatively obscure topic of DH research into historical resource.

While our analysis was conducted in Germany, its findings should apply to anybody working with digital research infrastructures of the PRC from within the legal provisions of the General Data Protection Regulation (GDPR) (European Parliament and Council, 2018). A systematic extension of our analysis to non-GDPR territories (USA) or a general comparison between MLSP2.0 and GDPR, is beyond the scope of this case study. Yet, anecdotal evidence and informal feedback suggest a mostly similar situation for all overseas territories. Digital research methods rely on digital infrastructures for the creation and distribution of digital artifacts. A discussion of the multifariousness of DH in a world of competing, and at times contradictory, legal frameworks for the application of intellectual property licenses, conflicting ethical guidelines related to data provenance, and global standard bodies dominated by Caucasian males with links to US technology companies is beyond the scope of this article.

Perhaps platforms dedicated to open access such as hugginface.co or indeed github.com can offer a different model for cooperation. All of this will require a more extensive discussion of the competing legal requirements placed on researchers, whether located in the PRC or the EU. Our research could be expanded in multiple directions. Firstly, it would be valuable to expand the global scope of our experiments towards other locations and subscription licenses, collating more data to better understand what CNKI users can expect to find and access from around the globe. It would secondly be interesting to test if Western platforms such as JSTOR or PubMed display similar behaviors when accessed from the PRC, or if there are differences between the US and the EU. Lastly, we should expand our view of research into Chinese Republican Newspapers, beyond the PRC, towards East Asia.

Funding Open Access funding enabled and organized by Projekt DEAL. This research has received partial funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 757365). Development of the Early Chinese Periodicals Online (ECPO) project was partially supported by the Excellence Initiative of the German Research Foundation (DFG), Cluster of Excellence "Asia and Europe in a Global Context", and the Heidelberg Centre for Transcultural Studies, Heidelberg University.

Data availability The data that support the findings of this study are available on Zenodo with the identifier https://doi.org/10.5281/zenodo.6801936. The secondary sources are documented in a Zotero library, https://www.zotero.org/groups/4046522/procedural_challenges/library.

Declarations

Competing interests The authors have no conflicts of interest to declare. The authors have no competing interests to declare that are relevant to the content of this article.

References

Duan, X 段晓林. (2016). Minguo wenxian shujuku kaifa xianzhuang yanjiu 民国文献数据库开发现状研究 (A Study on the Current Situation of Database Development of Republican Literature). 图书馆学研究 *Tushuguan xue yanjiu* (Researches on Library Science), *20*, 42–45. https://doi.org/10.15941/j.cnki.issn1001-0424.2016.20.007

EastView (2018). *China National Knowledge Infrastructure: China Academic Journals (CAJ)—All Series* (Version 2021-03-01). EastView. http://www.eastview.com/eastview_caj_allseries_title_list/. Accessed 27 May 2021.

European Parliament and Council (2018). *General Data Protection Regulation* (No. 2016/679). https://gdprinfo.eu/

FORCE11 (2014, October 9). *Guiding principles for findable, accessible, interoperable and re-usable data publishing version b1.0*. FORCE11. https://www.force11.org/fairprinciples. Accessed 26 May 2021.

Guo, W. 郭薇 & Ren, S. 任思琪 (2018). *Baozhi shuzi duixiang moxing sheji yu yingyong: Yi Shanghai tushuguan quanguo baokan suoyin pingtai wei li* 报纸数字对象模型设计与应用——以上海图书馆《全国报刊索引》平台为例(Design and Application of Newspapers Digital Object Model: Taking Quan Guo Bao Kan Suo Yin Platform of Shanghai Library for Example). 图书馆杂志 *Tushuguan zazhi* (Library Journal), *2018*(7), 41–52. https://doi.org/10.13663/j.cnki.lj.2018.07.006

Hansen, K. K., Buss, M., & Haahr, L. S. (2018). *A FAIRy tale*. https://doi.org/10.5281/zenodo.2248200

Li, M., 李明华 & Fang, C. 方丛蕙. (2019). *Tese guancang jianshe qudong xia minguo wenxian zhengli chuban qianjing yu celüe* 特色馆藏建设驱动下民国文献整理出版前景与策略研究 (research on the prospects and strategies for Organising and Publishing of Republican documents driven by the construction of Special Collections). 出版发行研究 chuban faxing yanjiu. *(Publishing Research)*, *4*, 90–94. https://doi.org/10.19393/j.cnki.cn11-1537/g2.2019.04.031.

Liebl, B., & Burghardt, M. (2020). *An evaluation of DNN architectures for page segmentation of historical newspapers*. https://arxiv.org/abs/2004.07317v1. Accessed 15 December 2020

Luo, Z. 罗振津 (2020). Tushuguan guancang zhenxi ziyuan qiangjiu yu baohu yanjiu: Yi tianjin tushuguan guancang jianguo qian zhongwen baozhi qiangjiu yu baohu wei li 图书馆馆藏珍稀资源抢救与保护研究——以天津图书馆馆藏建国前中文报纸抢救与保护为例 (Research on rescue and protection of rare resources in library: Taking the rescue and protection of Chinese Newspapers collected in Tianjin Library before the founding of the People's Republic of China as an Example). 图书馆工作与研究 *Tushuguan gongzuo yu yanjiu* (Library Work and Study), *10*, 89–93. https://doi.org/10.16384/j.cnki.lwas.2020.10.013

Oliveira, S. A., Seguin, B., & Kaplan, F. (2018). dhSegment: A generic deep-learning approach for document segmentation. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 7–12. https://doi.org/10.1109/ICFHR-2018.2018.00011

Peterson, D. A. M. (2020). Dear reviewer 2: go F' yourself. *Social Science Quarterly*, *101*(4), 1648–1652. https://doi.org/10.1111/ssqu.12824.

Smith, D. A., & Cordell, R. (2018). *Report: A research agenda for historical and multilingual Optical Character Recognition (OCR)* (p. 38). https://ocr.northeastern.edu/report/

Sturgeon, D. (2017). Unsupervised Extraction of Training Data for Pre-Modern, & Chinese, O. C. R. *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference*, 613–618. https://aaai.org/ocs/index.php/FLAIRS/FLAIRS17/paper/view/15490. Accessed 13 Nov 2019.

The Unicode Consortium (2020). *The Unicode Standard* (13.0). https://www.unicode.org/versions/Unicode13.0.0/UnicodeStandard-13.0.pdf. Accessed 28 May 2021.

Tiffert, G. D. (2019). Peering down the memory hole: censorship, digitization, and the fragility of our knowledge base. *The American Historical Review*, *124*(2), 550–568. https://doi.org/10.1093/ahr/rhz286.

Tsui, L. H., & Wang, H. (2020). Harvesting big biographical data for Chinese history: the China Biographical Database (CBDB). *Journal of Chinese History* 中國歷史學刊, *4*(2), 505–511. https://doi.org/10.1017/jch.2020.21

Tu, H. C., Hsiang, J., Hung, I. M., & Hu, C. (2020). DocuSky, a personal digital humanities platform for scholars. *Journal of Chinese History* 中國歷史學刊, *4*(2), 564–580. https://doi.org/10.1017/jch.2020.28

Wang, L. 王玲丽 (2014). Jindai wenxian suowei jiaojuan de shuzihua zhuanhua shijian—Yi Shanghai Tushuguan 'suowei jiaojuan shuzihua' xiangmu wei lie 近代文献缩微胶卷的数字化转化实践——以上海图书馆"缩微胶卷数字化"项目为例 (The practice of digital transformation of historical documents on microfilm: Taking the Shanghai Library's "Microfilm Digitization" Project as an Example). 图书馆杂志 *Tushuguan zazhi* (Library Journal), *3*, 52–55. https://doi.org/10.13663/j.cnki.lj.2014.03.009

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18.

Wu, F. 吴方枝. (2017). Shuzi renwen beijing xia minguo wenxian de shuzihua yanjiu 数字人文背景下民国文献的数字化研究 (a review of literature digitization of the Republican China Period in Digital Humanities). 图书馆学研究 *Tushuguan xue yanjiu (Research on Library Science), 15*, 18–21. https://doi.org/10.15941/j.cnki.issn1001-0424.2017.15.004

Wu, J., & Lam, O. (2017, March 9). *The evolution of China's Great Firewall: 21 years of censorship*. Hong Kong Free Press HKFP. https://hongkongfp.com/2017/09/03/evolution-chinas-great-firewall-21-years-censorship/. Accessed 29 May 2021.

Xia, C., & Bao, X. (2020). Dynamic digital humanities projects from Shanghai Library in China. In *Transformative Digital Humanities* (pp. 79–89). Routledge. https://doi.org/10.4324/9780429399923-9

Yang, J, 杨佳颖 & Xu, X. 许鑫 (2020). Minguo baozhi guanggao tuxiang ziyuan de yuyi biaozhu: Yi《xinwenbao》 suo kan de yueju guanggao wei li 民国报纸广告图像资源的语义标注——以《新闻报》所刊的越剧广告为例(Semantic Labeling of Advertising Images in Newspapers of the Republic of China Period: Illustrated by Shaoxing Opera Advertisements Published in Xinwen Bao). 图书馆杂志 *Tushuguan zazhi* (Library Journal), *online first*, 1–11. https://doi.org/10.13663/j.cnki.lj.2020.07.000

Yasuoka, K. (2019). Universal dependencies treebank of the four books in classical Chinese. *10th International Conference of Digital Archives and Digital Humanities*, 10. http://hdl.handle.net/2433/245217. Accessed 25 May 2021.

Zhang, X., & Xue, N. (2012). Extending and scaling up the Chinese Treebank Annotation. *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 27–34. https://aclanthology.org/W12-6306/. Accessed 13 Feb 2022.

Authors and Affiliations

Matthias Arnold[1] · **Duncan Paterson**[2] · **Jia Xie**[1]

Duncan Paterson
duncan.paterson@sinologie.uni-freiburg.de

Jia Xie
jia.xie@hcts.uni-heidelberg.de

[1] Heidelberg Centre of Transcultural Studies, University of Heidelberg, Heidelberg, Germany

[2] READCHINA, University of Freiburg, Freiburg im Breisgau, Germany