Journal of Open Humanities Data

The KIP Dataset: Socialist Popular Science in a Sino-Soviet Image Dataset Based on Chinese and Soviet Magazines (1956–1962)

DATA PAPER

YE ZHAO**
AZIM OSPANOV**
XUENAN CAO

*Author affiliations can be found in the back matter of this article
**Co-first authors

ubiquity press

ABSTRACT

The KIP (Knowledge Is Power) dataset compiles and transforms the visual content from 165 issues of the Chinese magazine *Zhishi Jiushi Liliang [Knowledge Is Power]* and the Soviet magazine of the same title, *Знание – сила [Knowledge Is Power]*, from 1956 to 1962. The captioned images are cropped from scans of the physical copies of the cover artworks (color), drawings (black-and-white), photographs (color and black-and-white), and artistic title lettering. The dataset contains 13,896 images (in PNG format) for historians of sciences and technology to study the shared sociotechnical imageries excavated from a transitional period in the Soviet bloc, where current archives on the same period are focused almost exclusively on textual data. The project is intended for researchers and scholars, and the KIP dataset is in no way a market substitute for the original magazines or their high-resolution digital copies.

CORRESPONDING AUTHOR:
Ye Zhao
Department of Communication, University of Macau, Macau
yc68013@connect.um.edu.mo

KEYWORDS:
China; Soviet; Popular Science; Magazine Arts; History of Science and Technology; Socialism

TO CITE THIS ARTICLE:
Zhao, Y., Ospanov, A., & Cao, X. (2026). The KIP Dataset: Socialist Popular Science in a Sino-Soviet Image Dataset Based on Chinese and Soviet Magazines (1956–1962). *Journal of Open Humanities Data*, 12: 148, pp. 1–6. DOI: https://doi.org/10.5334/johd.662

(1) OVERVIEW

REPOSITORY LOCATION

The KIP Dataset is deposited in the Harvard Dataverse. The dataset can be accessed at: https://doi.org/10.7910/DVN/KWUVPB. The Python scripts for image de-duplication using perceptual hashing (PHash with Hamming-distance thresholding) can be found at https://github.com/aziksh-ospanov/kip_dataset_scripts. Spectral image clustering scripts can be found at https://github.com/aziksh-ospanov/FKEA (Ospanov et al., 2024). The complete annotation set, along with the training configuration, inference scripts, and the fine-tuned weights, is available at https://github.com/elio11108/kip_annotation_scripts.

CONTEXT

This dataset is constructed from the research archive of Xuenan Cao's monograph, *Chinese Media Improvisations: Thriving on Deficits* (2026). The dataset is also linked to a collaborative computational humanities project on socialist popular science. The period from 1956 to 1962 is crucial for understanding information flows within the socialist world. The magazines used in creating this dataset were examples of the People's Republic of China's early popular science campaign, marked by the 1956 debut of the magazine *Zhishi Jiushi Liliang* (Knowledge Is Power), which includes translated articles from the Soviet magazines. Shared and reused imagery indicated by the overlaps of the Chinese and Soviet image data can be a useful starting point to investigate the specifics of Sino-Soviet information exchange during the early years of the Cold War. The two magazines copied each other's content without explicitly attributing the source. Although the unsourced overlaps cannot be used as evidence of any direct copy-and-pasting between these two magazines, the publication dates can offer interesting clues to historians of this period on the direction of knowledge transmission and information circulation.

Existing scholarship has examined the Soviet popular science (Andrews, 2003) and Chinese popular science publishing (Schmalzer, 2008) mostly in textual terms. In the Soviet Union, the popularization of science and technology was state-mandated, and trickled down into the communication systems (Korzun et al., 2024). The study offers some limited visual readings of *Знание – сила*'s illustrations and cover art. The visual layer within this transnational exchange remains understudied partly because the source materials are globally dispersed and not available in one aggregated dataset. In China, scholars have observed a translation boom of Soviet popular-science knowledge that took place from 1950 to 1961 (Liu, 2011; Y. Zhang, 2014), but there is scant knowledge about how much of that translation came with borrowed imagery. The KIP dataset intends to address these gaps by extracting and captioning the visual contents, including cover artworks, drawings, photographs, and artistic title lettering, from digitized bilateral issues over this seven-year period, and by exposing the results in a format that supports contemporary computational and historical analysis.

Existing collections of socialist visual materials include poster catalogues (van der Heijden et al., 2009) and page-level corpora of a single country's periodicals (*Pages of Early Soviet Performance*; Princeton Center for Digital Humanities & ITMO University, n.d.). Wong (2026) has released a textual corpus of a Chinese-language literary journal, *Chao Foon (Jiaofeng)*, published in Malaya from 1955 to 1961, the period KIP covers. Despite the same chronological window and the non-Euro-American context, Wong's OCR-driven corpus construction treats visual apparatus as ancillary to the scanned texts. The KIP dataset takes the opposite approach, curating the non-textual as primary evidentiary data and extending the paradigm from text to visual content in corpus building for periodical materials.

In terms of method and apparatus, the KIP dataset also contributes to a broader conversation in digital humanities about the construction of historical image corpora (Arnold & Tilton, 2019; Chelali et al., 2026) and the extension of digital humanities from text to image by computational analysis of digitized visual sources (Wevers & Smits, 2020). Furthermore, large-scale extraction efforts rely on training data in Latin scripts (Lee et al., 2020), and robust models for Asian-language layouts are costly to train (Shen et al., 2020). KIP takes a computational-lighter approach to non-Latin corpus construction.

(2) METHOD

DIGITIZATION & EXTRACTION

A total of 165 issues (1956–1962) of *Zhishi jiushi liliang* and *Знание – сила* are manually scanned. The PubLayNet Detectron models (Zhong et al., 2019) are used to extract figures from magazine pages by processing high-resolution digital scans of the magazines. Each scanned page is fed into a GPU-accelerated Detectron2 environment, run through LayoutParser v0.3.4, and configured with the PubLayNet (mask_rcnn_R_50_FPN_3x) figure-detection weights, which perform region proposal, region classification, and segmentation. We specifically segmented only the regions classified as "Figure" by the model, at a score threshold of 0.8 (full configuration in `README.md`, §10.1).

CLEANING

For data cleaning, we adopted a manual approach. We reviewed every crop produced by preliminary extraction and discarded crops that combined multiple images, retained residual text, or consisted of plain text misidentified as visual elements (artistic title lettering is kept). The cleaning was implemented for both the China and USSR portions. However, we found a significant asymmetry in the quantity and quality of figures between the two portions: the PubLayNet-trained detector under-segmented Chinese pages and was more susceptible to misclassifying blocks of Chinese plain text as figures than it does on Soviet pages. Therefore, we hand-annotated a small set of Sino-crops of mixed image-and-text regions sampled from Issues 1, 2, and 3 in 1956 of the Chinese magazine (N = 108 crops, 197 boxes, single class "image"). This annotation set was used to fine-tune a YOLOv8n detector. We applied the fine-tuned detector to preliminary extractions of the entire Chinese portion and re-extracted the pure-image regions. For quality control, each detection's confidence score was recorded in `kip_metadata.tab` (`det_conf_score`, N = 4,622 Chinese crops), and every resulting re-extracted crop was manually checked again. Full configuration is documented in the annotation repository.

STANDARDIZATION

Every figure was designated to the canonical form `kip_{country}_{yyyy}_i{NN}[_{section}]_p{PPP}_f{FF}.png`. The full specification, including the controlled section vocabulary (`cover`, `intercov`, `endcov`, `endpg`, `info`, `illus01–illus99`) and the `f{FF}` scoping rule, was parsed into the 14-column row schema in `kip_metadata.tab`, as a canonical filename for every single figure. Captions for 13,896 images were automatically generated by the BLIP-2 image-captioning model (Li et al., 2023; specific model weights: `Salesforce/blip2-opt-2.7b`), and follow the categorization rubric "*a {drawing|photo|cartoon|diagram} of* …" plus the most salient visual information. The captions, documented in `{China, USSR}_Caption Sheet.tab`, make the dataset easier to navigate and provide the detailed context future research needs to interpret each one meaningfully.

COMPARISON

To identify shared imagery, we used an adapted perceptual hash method for locating identical figures (image de-duplication; imagededup PHash, a 64-bit DCT hash). First, we resized images and converted them to greyscale to remove unnecessary detail. We then transformed images using the Discrete Cosine Transform (DCT) and encoded them into hash representations that capture each image's core visual structure (Ospanov et al., 2024). To identify duplicates, we compared these hash encodings using Hamming distance and removed near-duplicates. Hamming distance ≤ 10 (the default threshold of imagededup and of the released de-duplication script) flags near-duplicates; every cross-country candidate flagged was inspected visually and reported as cross-country duplicates (N = 53, 49 pairs and 4 triples) in `across_country_duplicates`. Spectral cluster summaries were produced by `FKEA_Evaluator.visualize_mode_by_eigenvectors_rff` on DINOv2 features (Oquab et al., 2024) and preserved in `spectral_clusters_by_years`.

QUALITY CONTROL

All quality control is steered manually and exhaustively. After each extraction round, every crop was checked as a pure image. This check first revealed under-segmentation in Sino materials

and prompted the annotation set. All 13,896 captions were likewise proofread, item by item, against the image for conformity to the standardization and accuracy.

(3) DATASET DESCRIPTION

REPOSITORY NAME

Harvard Dataverse (https://dataverse.harvard.edu)

OBJECT NAME

The KIP Dataset: Socialist Popular Science in a Sino-Soviet Image Dataset Based on Chinese and Soviet magazines (1956–1962).

FORMAT NAMES AND VERSIONS

PNG (images; distributed in 17 ZIP archives), CSV/TSV (tabular files), and Markdown (`README`). Each year-level archive unpacks to a folder of canonically named PNGs. The deposited `README.md` is the full documentation of record and instructions for the dataset (inventory §1; reconstruction §2; per-year counts §3; naming grammar §4; the 14 variables of metadata table §5; captions §6; derived layers §7; known limitations §8; provenance §9; pipeline and reproducibility §10; quality control §11; layer linkage §12; rights §13).

CREATION DATES

2025-01-01 to 2026-08-22

DATASET CREATORS

The dataset was conceived and supervised by Xuenan Cao; the figure extraction, segmentation, and substantial transformation were completed by Azim Ospanov; and the digitization, data curation, and all human-intervened steps were completed by Ye Zhao.

LANGUAGE

Chinese (Simplified, traditional pre-reform glyphs in earlier issues) & Russian; image captions are in English.

LICENSE

CC0 1.0 (for authors' own contributions; details see in `README.md`, §13)

PUBLICATION DATE

2026-08-26

(4) REUSE POTENTIAL

The KIP dataset can be reused for computational research in international and area studies, art and media histories, and histories of sciences. For historians of Sino-Soviet relations, the year-by-year distribution of cross-country duplicate clusters records when shared imagery surfaced in the two magazines; whether that chronology follows, departs from, or lags behind the technology transfer timeline documented by B. Zhang et al. (2006), and in which direction of the transmission flux, are illuminated for subsequent analysis. Analogous to how scientometrics reads cross-references as traces of knowledge diffusion (Blind & Fenton, 2022), the duplicate layer also furnishes shared images as traces of visual circulation.

The `{China, USSR}_Caption` datasheets support qualitative analyses of the Sino-Soviet exchange, especially by annotating the socialist cosmic imaginaries of that compressed development era and its socialist cosmopolitanism, the early PRC's participation in a socialist

transnational culture instead of a retreat from worldliness (Volland, 2017; Cao, 2026). For historians specializing in visual and fine arts, it can help capture temporal image themes crystallized with rife Communist sentiment (Sankova & Moscow Design Museum, 2020).

DATA ACCESSIBILITY STATEMENT

The KIP dataset is openly available in the Harvard Dataverse at https://doi.org/10.7910/DVN/KWUVPB; the scripts, annotation set, and fine-tuned weights are available from the GitHub repositories listed under Repository Location.

ACKNOWLEDGEMENTS

The authors thank Qinqin Zhang and Jack Ng, the research data librarians at The Chinese University of Hong Kong (CUHK). The authors also thank Dr. Vladimir Lukin, Postdoctoral Associate at Duke University.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Ye Zhao: Data Curation, Project Administration, Methodology, Resources, Validation, Writing.

Azim Ospanov: Data Curation, Investigation, Methodology, Formal Analysis, Validation, Writing.

Xuenan Cao: Conceptualization, Methodology, Project Administration, Resources, Supervision, Writing.

AUTHOR AFFILIATIONS

Ye Zhao orcid.org/0009-0003-5815-8070
Department of Communication, University of Macau, Macau

Azim Ospanov
Department of Computer Science & Engineering, The Chinese University of Hong Kong, Hong Kong

Xuenan Cao orcid.org/0000-0001-7830-8708
General Education (Science, Technology, and Society), San Francisco Bay University, USA

REFERENCES

Andrews, J. T. (2003). *Science for the masses: The Bolshevik state, public science, and the popular imagination in Soviet Russia, 1917–1934*. College Station: Texas A&M University Press.

Arnold, T., & **Tilton, L.** (2019). Distant viewing: Analyzing large visual corpora. *Digital Scholarship in the Humanities*, *34*(Supplement 1), i3–i16. https://doi.org/10.1093/llc/fqz013

Blind, K., & **Fenton, A.** (2022). Standard-relevant publications: Evidence, processes and influencing factors. *Scientometrics*, *127*(1), 577–602. https://doi.org/10.1007/s11192-021-04210-8

Cao, X. (2026). *Chinese media improvisations: Thriving on deficits*. Stanford, CA: Stanford University Press. https://doi.org/10.1515/9781503647244

Chelali, M., Gosselet, S.-K., Cloppet, F., Kurtz, C., Bloch, I., & **Foliard, D.** (2026). The FORBIN dataset: A collection of historical photographs with archival metadata. *Journal of Open Humanities Data*, *12*, 56. https://doi.org/10.5334/johd.487

Korzun, V. P. [Корзун, В. П.], Limanova, S. A. [Лиманова, С. А.], & **Nosova, M. S. [Носова, М. С.].** (2024). Популяризация научного знания в СССР в 1945–1975 гг.: между социальным заказом и академической культурой [Popularization of scientific knowledge in the USSR, 1945–1975: Between social order and academic culture]. *Russianstudies.Hu*, *6*(2), 115–132. https://doi.org/10.38210/RUSTUDH.2024.6.18

Lee, B. C. G., Mears, J., Jakeway, E., Ferriter, M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K., & **Weld, D. S.** (2020). The Newspaper Navigator dataset: Extracting headlines and visual content from 16 million historic newspaper pages in Chronicling America. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 3055–3062). https://doi.org/10.1145/3340531.3412767

Li, J., Li, D., Savarese, S., & **Hoi, S.** (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (PMLR 202, pp. 19730–19742). https://proceedings.mlr.press/v202/li23q.html

Liu, X. [刘新芳]. (2011). 中华全国科学技术普及协会科普工作史评 [A historical review of the popularisation work of the National Association for the Dissemination of Science and Technology]. *科普研究* [*Studies on Science Popularization*], *6*(4), 50–56.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., … **Bojanowski, P.** (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*. https://arxiv.org/abs/2304.07193

Ospanov, A., Zhang, J., Jalali, M., Cao, X., Bogdanov, A., & **Farnia, F.** (2024). Towards a scalable reference-free evaluation of generative models. In *Advances in Neural Information Processing Systems 37* (NeurIPS 2024). https://arxiv.org/abs/2407.02961

Princeton Center for Digital Humanities, & **ITMO University.** (n.d.). *Pages of Early Soviet Performance (PESP)* [Data set]. Hugging Face. https://huggingface.co/datasets/ajanco/pesp

Sankova, A., & **Moscow Design Museum.** (2020). *Soviet space graphics: Cosmic visions from the USSR*. Phaidon.

Schmalzer, S. (2008). *The people's Peking Man: Popular science and human identity in twentieth-century China*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226738611.001.0001

Shen, Z., Zhang, K., & **Dell, M.** (2020). A large dataset of historical Japanese documents with complex layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 548–549). https://arxiv.org/abs/2004.08686

van der Heijden, M., Landsberger, S. R., & **Shen, K.** (2009). *Chinese posters: The IISH-Landsberger collections*. Prestel.

Volland, N. (2017). *Socialist cosmopolitanism: The Chinese literary universe, 1945–1965*. Columbia University Press. https://doi.org/10.7312/voll18310

Wevers, M., & **Smits, T.** (2020). The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities*, *35*(1), 194–207. https://doi.org/10.1093/llc/fqy085

Wong, N. Y. H. (2026). Modernism's vector: Dataset for a Chinese-language journal in late 1950s Malaya. *Journal of Open Humanities Data*, *12*, 40. https://doi.org/10.5334/johd.485

Zhang, B., Zhang, J., & **Yao, F.** (2006). Technology transfer from the Soviet Union to the People's Republic of China: 1949–1966. *Comparative Technology Transfer and Society*, *4*(2), 105–167. https://doi.org/10.1353/ctt.2006.0024

Zhang, Y. [张亚娜]. (2014). 苏联科普作品对中国科普发展影响研究 [A study of the influence of Soviet popular-science works on the development of science popularisation in China]. *科普研究* [*Studies on Science Popularization*], *9*(4), 72–77.

Zhong, X., Tang, J., & **Jimeno Yepes, A.** (2019). PubLayNet: Largest dataset ever for document layout analysis. In *Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1015–1022). https://doi.org/10.1109/ICDAR.2019.00166

TO CITE THIS ARTICLE:
Zhao, Y., Ospanov, A., & Cao, X. (2026). The KIP Dataset: Socialist Popular Science in a Sino-Soviet Image Dataset Based on Chinese and Soviet Magazines (1956–1962). *Journal of Open Humanities Data*, 12: 148, pp. 1–6. DOI: https://doi.org/10.5334/johd.662

Submitted: 15 September 2026
Accepted: 21 October 2026
Published: 13 November 2026

COPYRIGHT:
© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See https://creativecommons.org/licenses/by/4.0/.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.