Supplementary methodology

for

A Web-based Database on Exposure to Persistent Organic Pollutants in China

Zhaomin Dong[1,2], Xiarui Fan[1], Yao Li[1], Ziwei Wang[1], Lili Chen[3], Ying Wang[1,2], Xiaoli Zhao[4*], Wenhong Fan[1,2*] and FengChang Wu[4]

[1,] School of Space and Environment, Beihang University, Beijing, China
[2,] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing, China
[3,] Beijing Academy of Edge Computing, BAEC, Innovation Building, HaiDian district, Beijing, 100191, China
[4,] State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing 100012, China

*Corresponding authors:
Correspondence to Prof. Xiaoli Zhao, State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing 100012, China. zhaoxiaoli_zxl@126.com.

Correspondence to Prof. Wenhong Fan, School of Space and Environment, Beihang University, Beijing, 100191, China. fanwh@buaa.edu.cn.

Study framework. As illustrated in **Figure 1**, the process of establishing a web-based database for POPs consisted of five steps. In step 1, we created the chemical panels that comprise all of the POPs listed in the Stockholm Convention (http://www.pops.int/TheConvention/ThePOPs/AllPOPs/tabid/2509/Default.aspx). In step 2, we identified all available literature from four databases, including the PubMed, web of science, Scopus and National Knowledge Infrastructure (CNKI). In step 3, we excluded the irrelevant literature based on title, abstract and full-text examination, and manually extracted the monitoring data from all candidate literature. In addition, the literature quality was also evaluated. Next in step 4, we performed spatiotemporal analysis and basic risk assessment of exposure data harmonized in step 3. Finally, the publicly available online database was established.

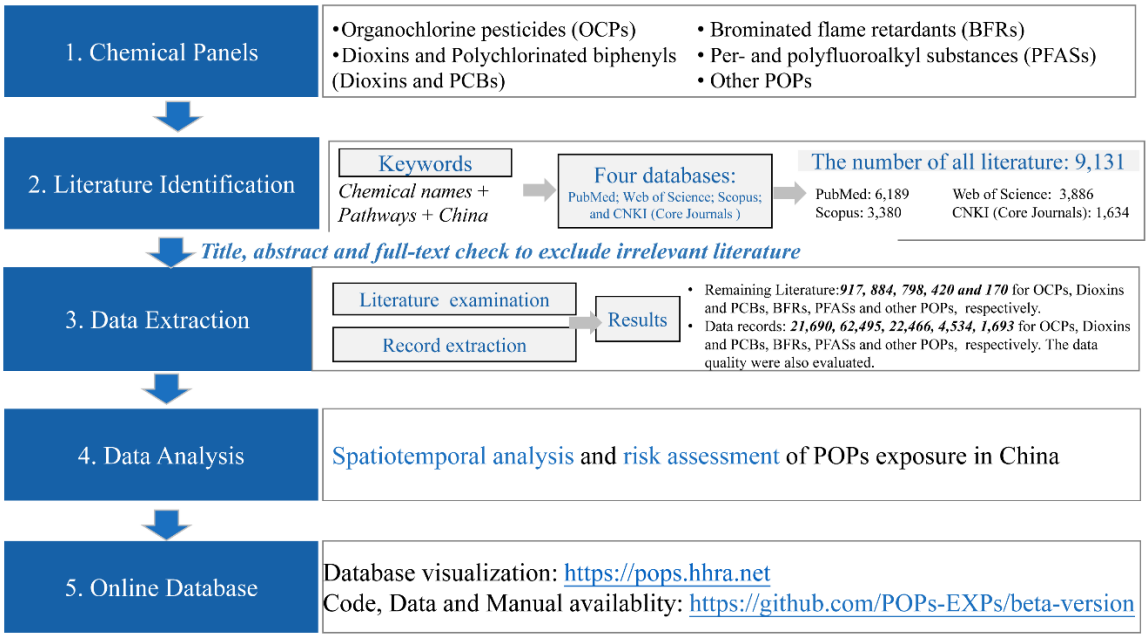

Figure 1. The framework used to establish online database of exposure to persistent organic pollutants (POPs) in China.

Chemical panels. In this study, the target POPs were divided into the following groups: 1) organochlorine pesticides (OCPs), comprising aldrin, chlordane, chlordecone, dichlorodiphenyltrichloroethane (DDT), dicofol, dieldrin, endrin, endosulfan, hexachlorobenzene (HCB), hexachlorocyclohexane (HCH), heptachlor, mirex, pentachlorobenzene, pentachlorophenol (PCP) and toxaphene; 2) dioxins and polychlorinated

biphenyls (dioxins and PCBs), comprising polychlorinated dibenzo-*p*-dioxins (PCDDs), polychlorinated dibenzofurans (PCDFs) and PCBs; 3) brominated flame retardants (BFRs), comprising polybrominated diphenyl ethers (PBDEs), hexabromocyclododecane and hexabromobiphenyl; 4) per- and polyfluoroalkyl substances (PFASs), comprising perfluorooctane sulfonate (PFOS) and perfluorooctanoic acid (PFOA); and 5) other POPs, comprising short-chain chlorinated paraffins (SCCPs), hexachlorobutadiene (HCBD) and polychlorinated naphthalenes (PCNs).

Literature identification. To retrieve relevant papers, a systematic search was performed with a query based on the combined keywords of "*chemical names*" and "*exposure pathways*" and "*China.*" Multiple chemical names were used for each chemical to retrieve as much of the available literature as possible. We also denoted 11 pathways of exposure, as follows: air, soil, dust, sediment, water, food, biological, blood, serum, plasma and breast milk.

As mentioned above, we identified the literature from four databases. Regarding on the PubMed database, we utilized a R-based web crawler (the code termed as *code_literature_search.zip* has been already uploaded to the https://github.com/POPs-EXPs/beta-version) to identify target literature, with the combined keywords for search were defined in the code. The query results obtained from the web crawler were downloaded in Extensible Markup Language format, and the titles, authors, publication dates, journals, PubMed IDs (PMIDs) and DOIs of the resulting publications were extracted by text mining in R.

With respect to the web of science and Scopus database, we used the following term to search relevant reports:

(DDT or Dichlorodiphenyltrichloroethane or Aldrin or Chlordane or Chlordecone or PBDE or Polybrominated Diphenyl Ethers or Hexabromodiphenyl ether or HexaBDE or Hexa-BDE or heptabromodiphenyl ether or Decabromodiphenyl ether or Tetrabromodiphenyl ether or heptaBDE or DecaBDE or TetraBDE or hepta-BDE or Deca-BDE or Tetra-BDE or Dicofol or Dieldrin or Endrin or Heptachlor or Hexabromobiphenyl or Hexabromocyclododecane or Hexachlorobenzene or Pentachlorobenzene or Hexachlorobutadiene or alphahexachlorocyclohexane or alpha-HCH or betahexachlorocyclohexane or beta-HCH or

Lindane or hexachlorocyclohexane or HCH or Mirex or Pentachlorophenol or PCDD or polychlorinated dibenzo-p dioxins or Polychlorinated dibenzodioxins or PCDF or Polychlorinated dibenzofurans or PCDD/Fs or Polychlorinated biphenyls or PCBs or Polychlorinated naphthalene or PCN or PCNs or PFOA Perfluorooctanoic acid or Perfluorooctanesulfonic acid or PFOS or short chain chlorinated paraffins or SCCPs or short-chain chlorinated paraffins or endosulfan or Toxaphene) **And** (*air or atmosphere or soil or dust or water or sediment or plant or food or intake or uptake or blood or serum or plasma or breast milk*) **And** China.

Similarly, we identified the research in the CNKI database (Core journals: *https://kns.cnki.net/KNS8/AdvSearch?dbcode=CJFQ*) by following the query sentence (**Note:** we directly expressed **the query sentence in Chinese** because we attempted to identify the reports written in Chinese):

SU=('二恶英' + '二噁英'+ 'dioxin' + '多氯联苯' + 'PCB' + '多溴联苯醚' + 'PBDE'+ 'HBCD' + '六溴代二苯' + 'HBB' +'全氟化合物' + 'PFASs' + 'PFCs' + '氯化石蜡' + 'SCCPs' + '毒杀芬'+ 'Toxaphene' + '六氯丁二烯' + 'Hexachlorobutadiene' + 'polychlorinated naphthalene' + '多氯萘'+ '滴滴涕' + 'DDT'+ '六六六' + 'HCH'+ 'Aldrin' + '艾氏剂'+ 'Chlordane' + '氯丹'+ 'Chlordecone'+ '十氯酮'+ 'Dicofol' + '三氟杀螨醇'+'Dieldrin'+ '狄氏剂'+'Endrin'+ '异狄氏剂'+'Endosulfan' + '硫丹'+ 'Hexachlorobenzene' + '六氯苯'+ 'Hexachlorocyclohexane' + '六氯环己烷'+'Heptachlor' +'七氯'+'Mirex'+ '灭蚁灵'+ 'Pentachlorobenzene' + '五氯苯'+'Pentachlorophenol' + '五氯苯酚') **And** SU=('大气' +'水'+'土壤' +'沉积物'+'膳食'+'食物'+'粉尘'+'血液'+'尿液'+'乳汁'+'头发').

After the remove of duplicate records and initial examination on irrelevant literature, the final set of literature was compiled into a list for further analysis, as shown in *lit_combine.xlsx* in the *code_literature_search.zip.*

Data extraction, analysis and quality evaluation. The publications that were identified by the above steps were first downloaded and curated in terms of titles, abstracts and full texts. We precluded literature that did not report monitoring data in China, such as toxicity studies, model simulations, laboratory experiments, emissions studies, or others. Then, the data from the remaining literature were manually extracted and cross-checked by two authors. The

monitoring results for the associated congeners of PBDEs, PCBs, PCDDs, PCDFs and some OCPs were also recorded. For each record, we included the following available items: chemical panel, chemical name, congener, PMID, sample location (province, city and site), sample time (year and season; if the year of sampling was not available, we used the publication year as alternative and marked it as a note), sample size, pathway, unit and statistics (range, mean, median and standard deviation) and method detection limit. All of the monitoring results are provided in the folder *exposure_data* in the Github (https://github.com/POPs-EXPs/beta-version).

Then, the extracted data were harmonized by data cleaning, unit conversion and other treatments. The resulting harmonized data were then ready for further examination, such as spatiotemporal analysis and risk assessment in step 4.

To evaluate the quality of extracted data, we also obtained the information on 'detection limit', 'instrumental methodology', 'recovery rate', 'internal standard', 'sample size', 'sampling location' for research articles. Then, we established a scoring matrix based on quality assurance/quality control and sample size:

Table 1. The scoring matrix for data quality

Factors		Score
Detection limit	Reported	2
	Not reported	1
Internal standards	Reported	2
	Not reported	1
Recovery rate	>50%	3
	≤50%	2
	Not reported	1
Sampling size	>50	3
	20-50	2
	<20 or Not reported	1
Total		10

After that, we rated the literature and associated data based on the total score (**Table 2**)

Table 2. The classification of data quality based on total score

Total score	Data quality
9-10	High
6-8	Medium
≤5	Poor

Online database, code and data availability. In Step 5, the online database and data visualization were developed using the R package *Shiny*, which is capable of building interactive web applications directly from R. Then, we registered our website at https://pops.hhra.net to display the online database, and the framework of the website (https://pops.hhra.net) is illustrated in **Figure 2.** This project was also posted on GitHub, and users can download the code (https://github.com/POPs-EXPs/beta-version) and run the main document termed "*app.R*" in R to access the database tool on local computers.

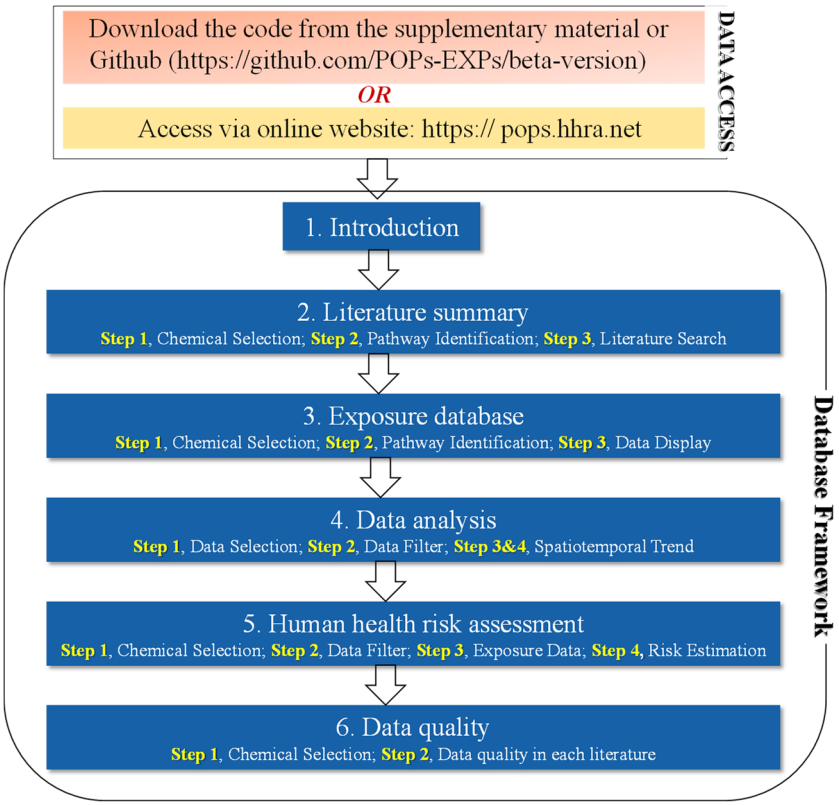

Figure 2. The web design for the database on exposure to persistent organic pollutants in China (https://pops.hhra.net).

The sustainability of database. To certain the sustainability, the database will be annually updated. Particularly, we will first identify the latest publications in the early June of the calendar year, and then complete the data extraction by the end of July. After that, we anticipate to upload all data to website by the mid-August of each calendar year.