Data Management and Analysis Core [1]

The objective of this core is to apply informatics to support and optimize the ISRP scientific research process, training, and methods to maximize research outcomes, applied solutions, replicable products, and sound evidence-based decision support. The core provides expert staff, platforms, services, and research support integrating five Aims:

**Aim 1:** Develop, maintain, and automate data management, data sharing, and quality assurance infrastructure for full reproducibility, transparency, and rigor in all ISRP studies.

The Data Management and Analysis Core (DMAC) and Analytical Core [2](AC) meet jointly on a weekly basis. During year 1 the data management team led discussions on topics including: tabular data structure; data authorship guidelines by the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS); and leveraging repository systems and FAIR practices to increase the impact of research data. The DMAC team presented to the monthly ISRP meeting on ‘tidy’ tabular data structure, creation and utility of data dictionaries, and file organization best practices.

The DMAC supported the publication of data in subject-specific data repositories where they exist for specific types of data. For example, some environmental data sets have been deposited in Pangaea, which provides a curation service to assist with conversion of tabular data to a structure unique to that repository. The DMAC is also working with ISRP teams on retroactive publishing of data from past projects, such as the AESOP project.
The DMAC has begun to meet with individual project teams to learn more about their data management and sharing practices and needs, including data-specific metadata and publishing options. This work began with several of the trainees and their data publications, and based on this information gathering, we are providing instruction and resources that will improve data management and integration within the program, and for external users and the public. In parallel, we will collaborate with the project teams to review and refine their data management plans as needed to fully support robust data sharing, reuse, and reproducibility.

Aim 2: Support ISRP Projects and Cores with embedded expert biostatistical contributions, services, and guidance.

During year 1, the DMAC was very busy assisting with data analysis. The statistics team led a discussion on measures of similarities of two PCB congener profiles and ways to test for such similarities. The statistics team provided guidance and analysis of data on several studies concerning PCB congeners in paints, tissues, and indoor air. The weekly AC/DMAC meeting also provides a platform for discussing statistical techniques pertinent to PCB data analysis and statistical issues in student presentations.

Working with AC members, the DMAC team worked out a protocol governing the PCB data submitted to DMAC for analysis, which covers the variable names in the Excel file and the convention for file naming. An online form for submitting priority analysis plans to DMAC for data analysis has been implemented. It is implemented using the Qualtrics Survey Software and is in its trial phase.

Aim 3: Develop novel statistical methods and associated software for data analytic challenges that impact all ISRP Projects and Cor
DMAC serves the ISRP by also developing methods for the shared issue of profile analysis of complex toxicological and metabolomic concentration profiles. DMAC will develop and refine methods for estimating the covariance matrix of congener measurements using tens of thousands of samples analyzed by the Analytical Core to date and all machine-readable samples that can be compiled from public repositories.

**Aim 4: Support the Research Experience and Training Coordination Core (RETCC)** [3] by providing guidance, resources, events, and instruction on data science and informatics to trainees and investigators.

During year 1, the DMAC became part of the RETCC [3] advisory committee and provided regular input on courses and instruction on data management and analysis for trainees and faculty. The Data Services Librarian and Engineering Librarian co-taught a one-credit course, “Managing Data to Facilitate Your Research,” via the Civil and Environmental Engineering Department, CEE: 5110, during Spring semester 2020. The course will be taught again in the Spring semester 2021, cross-listed with Occupational and Environmental Health (OEH) in the College of Public Health to broaden our outreach to more potential students in the ISRP, and other STEM graduate students.

**Aim 5: Provide the integrative data management and analytical foundations for ISRP-wide efforts to quantify, constrain, and communicate uncertainties in estimating and projecting PCB exposomes of the U.S. school-age population and currently available means for reducing them.**
During the first year DMAC led ISRP-wide discussions of data-driven research integration and developed study design for a systematic review on the parametric ranges and uncertainties in outdoor, home, school, dietary exposures for children in the United States. We began to build and publish FAIR modeling infrastructure for emissions, dispersion, and exposures in outdoor air from U.S. and global datasets. A first publication on this topic is in review, with application to PCB emissions from e-waste processing in India.

Core Leader: Kai Wang, PhD

Dr. Wang has served as Biostatistician for the ISRP since its inception in 2006, and has extensive experience in analyzing data arising from ISRP projects. Together with Co-Leader Spak, he will be responsible for day to day management and direction of the DMAC with a particular focus on data analysis aims.

Co-Core Leader: Scott Spak, PhD

Dr. Spak is an Assistant Professor of Urban and Regional Planning, Civil and Environmental Engineering, and Environmental Policy. Dr. Spak has 15 years of experience in the modeling and analysis of POPs policies, emissions, and chemical transport. Together with Wang, Spak will be responsible for day-to-day management and direction of the DMAC, focusing on data management, data sharing, and integration objectives. He will lead software development and implementation for Aim 1 data management and data sharing; direct Aim 3 profile dataset compilation and collaborative investigation with ISRP and the UI3 Working Group; and lead Aim 5 integration, cross-center collaboration, and software development for research and decision support.

Co-Investigator: Michael Jones, PhD

Dr. Jones is a Professor of Biostatistics in the College of Public Health. He is an established biostatistician and has served as a member of the ISRP team since 2014. He has extensive experience in data analysis in general and left-censoring data analysis in particular. Jones will assist Project 3. Working with Wang, he will provide support for Projects 1, 2, and 5, develop new statistical methods for analyzing congener measurements when some of them are below detection limits, and R programing to implement these methods.

Data Services Manager: Brian Westra, MS

Brian is the UI Libraries Data Services Manager. He leads development of institutional data services, infrastructure, and policies. He has 20 years of experience in data management and data services development and implementation. Westra will lead data management and quality assurance infrastructure implementation and operations; serve as primary DMAC lead for development and implementation of data management plans for ISRP projects and cores; and conduct and support training activities.

Data Specialist, Qianjin (Marina) ZHang

Marina is the Engineering & Information Librarian at Lichtenberger Engineering Library, and leads data management education for UI3, the College of Engineering, and the Department of Computer Science. Zhang will lead and coordinate Aim 3 instruction and training activities, serve as primary DMAC lead for trainee Individual Development Plans, and support Aim 1 data management, data sharing, and quality assurance activities.