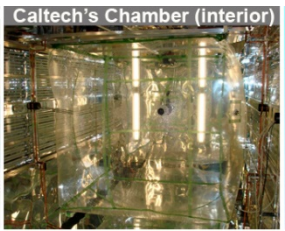
What is realistic and doable for an atmospheric chemistry database?

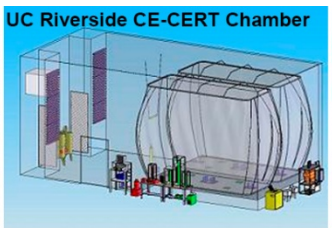
Tran B. Nguyen

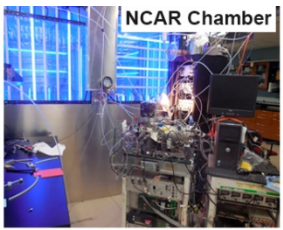
University of California, Davis

ICARUS (Index of Chamber Atmospheric Research in the United States)

- Open-access database development project for atmospheric chamber studies
  - Motivation: archive 10+ years of data and streamline future data submissions
  - Initial cohort of 13 research groups, will be open to all
  - Data management guidance from NCAR/DSET







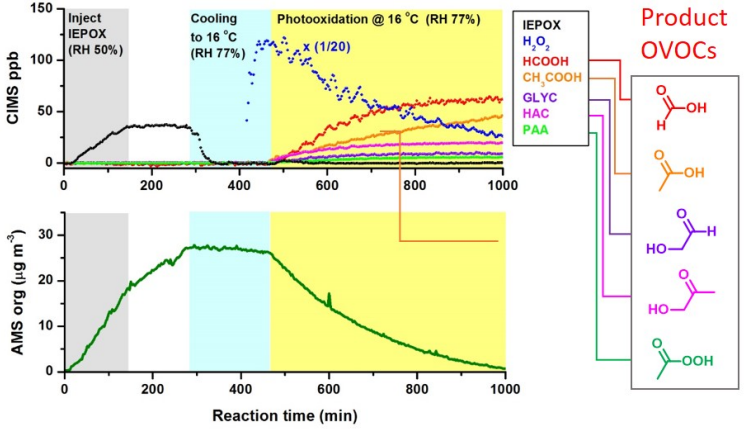
Considerations for our domain-specific database

We have the typical challenges:

- Need to deal with heterogeneous & non-digital data, get consensus on metadata/data standards, etc…

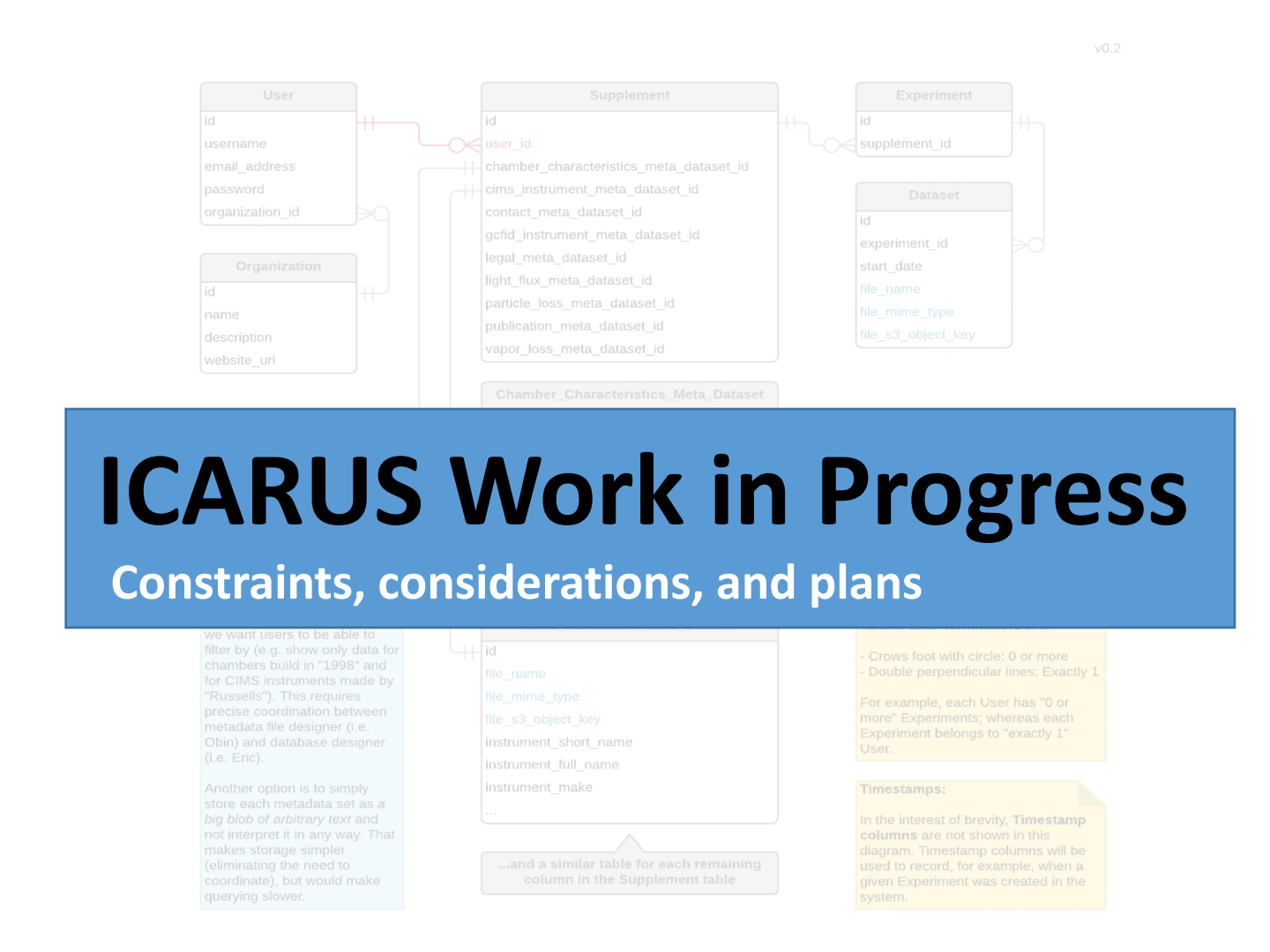
With some key features:

- The community has high volumes of legacy data & no consistent data management protocols
- Unlike observational data, each experiment needs a detailed road map
  - High quantities of metadata needed



Data accessibility & discoverability does not translate to data reuse

- European counterpart to ICARUS: EUROCHAMP 1,2, and 2020
  - The 1&2 databases saw very limited use
    - Out of ~1000 experiments, ~900 experiments have never been viewed. Only ~10 have been used in some way
    - **Reasons:**
      - Lack of supporting information (metadata)
      - No details on how to correct for chamber specific effects (“wall losses”)
      - Lack of publicity
      - Too many experiments with low value to end-users

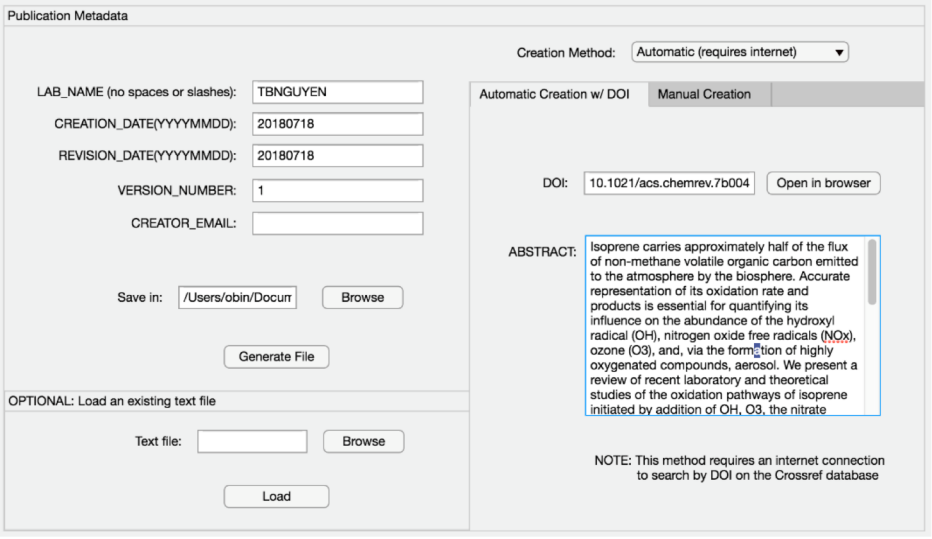


ICARUS Work in Progress

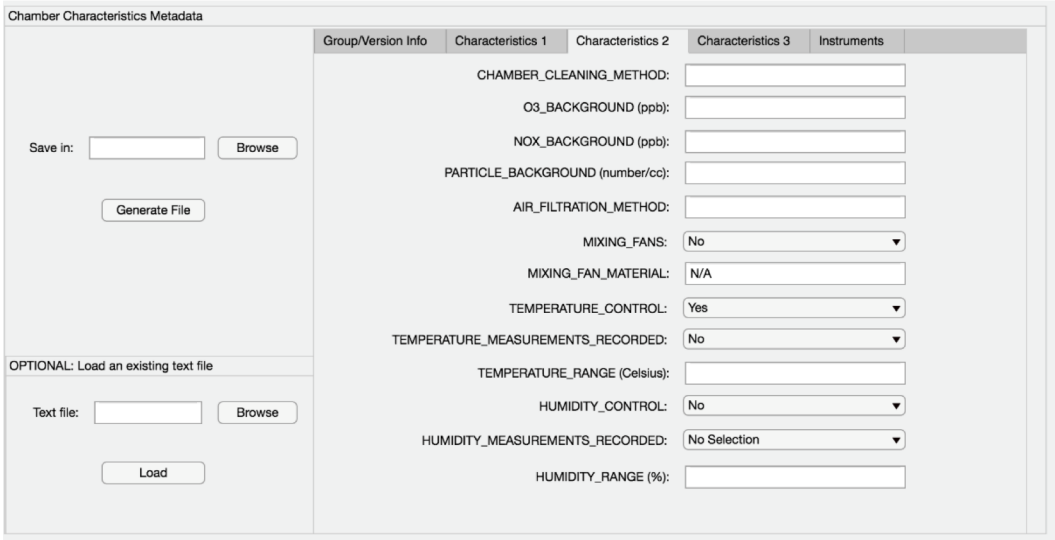
Constraints, considerations, and plans

Developing tools to homogenize data

**use metadata from preexisting database, e.g., crossref*



**accepts file template to prepopulate fields*



“Experimental metadata” is the bottleneck, especially for legacy data

RECORD_ID= JSEIN20140123
ISO_ASSET_TYPE=Experiment
RESOURCE_TYPE= Experiment Metadata
EXPERIMENT_CATEGORY= Isoprene ozonolysis series
EXPERIMENT_TITLE=Isoprene ozonolysis under humid conditions (no scavenger)
EXPERIMENT_DATE (YYYYMMDD)=20140123
VOC_NAME=Isoprene
VOC_INITIAL_CONC (ppb)=100
EXPERIMENT_RH (%)=51
EXPERIMENT_T (deg C)=25
SEEDED_EXPERIMENT= No
TYPE_OF_SEED= N/A
SEED_INITIAL_CONC (ug/m^3) = N/A
REACTION_TYPE= Dark reaction
OXIDANT_NAME= Ozone
OXIDANT_INITIAL_CONC (ppb)= 600
RO2_MAIN_FATE= HO2 {controlled: HO2, NO, NO2, RO2, NO3, isomerization, loss}
RO2_LIFETIME (sec) = 0.1

EXPERIMENT_GOALS=This experiment is designed to be easily compared to our dry no-scavenger isoprene ozonolysis experiment (on January 6th), with the only change in conditions being the humidity (50% RH instead of 3-4%). Our main goal is to see how the products of humid ozonolysis compare with those of dry ozonolysis. The hot, wet conditions at the end may also help with ongoing investigation of GTHOS interference.
EXPERIMENT_SUMMARY=The experiment went as planned; initial conditions all matched the values we were aiming for, with 601 ppbv of ozone, 53% RH, and approximately 100 ppbv of isoprene at 25 degrees C. Ozonolysis proceeded rapidly, with the CF3O- CIMS observing the usual ozonolysis products (e.g. HMHP), and as we did not use an OH scavenger, the CIMS saw evidence of ISOPOOH and IEPOX formation as well. At the end, the temperature was ramped to 45C, during which a number of ozonolysis product signals increased and GTHOS was able to observe the temperature-dependence of interferences
EXPERIMENT_LOCAL_STARTTIME (hh:mm)=20:00

EXPERIMENT_TIMELINE=
15:05: T/RH/NOx/O3 sampling; bag cooled to 25C.
15:30: Began humidification.
17:06: Stopped humidifying; continued filling bag with dry air.
17:49: Stopped filling bag.
17:56: Started O3 injection.
19:30: Started adding air to refill and dilute bag.
19:38: Stopped O3 injection.
19:42: Stopped dry air injection.
20:00: Injected 10.25 uL isoprene.
20:19: Stopped injection.
25:00: Set temperature to ramp to 45C.
25:40: Temperature stabilized in the chamber (still appeared to be ramping in bag).
26:10: Temperature fully stabilized in bag.
26:25: Bag cooling and flushing.

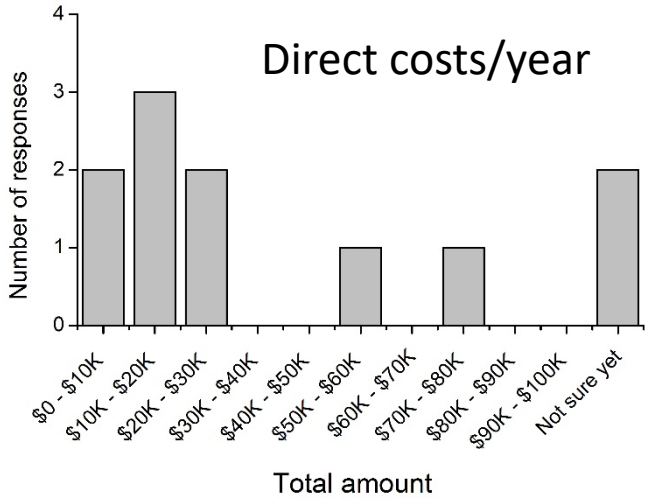
- What we’re doing differently from other databases in the field (and related fields):
  - We require a consistent naming scheme and format (form-generated)
  - All documents will be machine-readable with tools to read into Matlab and other data processing programs
  - We want to influence research practices in data management!

Legacy data: each group has different data volumes and funding needs

- Before database project started, we asked: "How much financial support do you need to do **initial archiving** and **development of SOP**?"

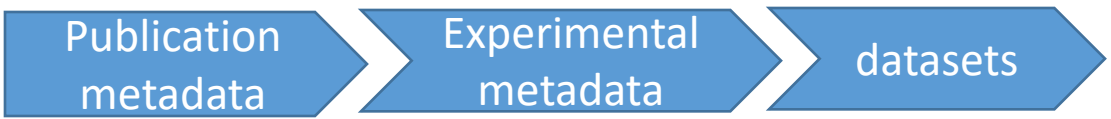
*"The needed financial support I list is a guess, **since we have never done this before**. We have about 15 years of data of different types, and I am not sure of everything that is available."*

*"We need to figure out what the **ultimate objective of this data** is and how it will be used. This will help prioritize the information to be included, the format needed, etc., which in turn will drive the actual cost of initial data summary (as will the number of years to go back…"*



Legacy data: Priorities and compromises

- **We can't (and shouldn't?) archive everything**
  - Priority scale:
    1. Published chamber data with high quantities of metadata;
    2. Published chamber data with low quantities of metadata;
    3. Unpublished data with high quantities of metadata; and
    4. Unpublished data with lower quantities of metadata



- Three groups (out of 13) deemed digitizing experimental notes for legacy data too cumbersome of a task
  - These groups will provide experimental details ***moving forward*** (this affected the funding they requested)

Addressing **data quality**

- Tricky to judge “quality” (not something we want to do)
  - Is one dataset as good as another?
  - How to reconcile discrepancies?
  - Are differences in results due to differences in analytical methods or chamber operation?

- Aim to provide the users with as much information as possible to judge for themselves

- As researchers, we plan to carry out inter-comparison campaign to map out chamber performance
  - Perform standard experiments identically with overlapping equipment to rule out some sources of error

Future plans

- Forge onward with database and tool development, upload test datasets
  - Integrate some great ideas from this workshop!
- Beta testers will give feedback after test uploads
  - Revise data model as needed
- Integrate with the Digital Assets Services Hub (DASH) at NCAR for long-term data management
  - Removes need for sustained funding

Acknowledgments

- **ICARUS Steering Committee members + their group members**
  - D.R. Cocker, W.P.L. Carter, N.M. Donahue, A.L. Robinson, A.P. Kaduwela, L. Hildebrandt-Ruiz, J.-L. Jimenez, N.L. Ng, S.A. Nizkorodov, J.H. Seinfeld, S.N. Pandis, G.S. Tyndall, J.J. Orlando, P.J. Ziemann

- **UC Davis Staff**
  - P.O. Sturm, E.E. Cavanna

- **NCAR/DSET collaborators**
  - S.J. Worley, M.S. Mayernik, S. Hou

- **NSF/GEO/AGS funding**
  - Collaborative grants AGS-1740571, AGS-1740587, AGS-1740665, AGS-1740640, AGS-1740568, AGS-1740552, AGS-1740610, and AGS-1740625.