How to extract duplicate references using R script and data cleaning in Excel

KATIE PEARSON – DECEMBER 1, 2020

DEVELOPED FOR THE CALIFORNIA PHENOLOGY NETWORK

This code was developed to rapidly georeference herbarium specimens by importing georeference data that already exist for duplicate specimens. The code looks through each record in a provided dataset and determines whether that record is found in the *omoccurduplicatelink* table (i.e., a duplicate has been linked to that record in your Symbiota portal). If a duplicate is found, it determines whether the duplicate is georeferenced. If none of the duplicates are georeferenced, the code moves to the next specimen in the provided dataset. If a georeferenced duplicate is found, the code will add these data into a new output file (newMycoll) such that the unique identifier (occid) corresponds to the occid in the **input** dataset and the georeference data is copied from the duplicate record. The user must then clean the output file so it results in one row per occid/specimen record).

INPUT:

- Dataset (CSV format) from target collection containing, at minimum, the columns: **occid** (unique identifier), **catalogNumber**, **otherCatalogNumber**, **collectorNumber**, **collector**, **decimalLatitude**, **decimalLongitude**, **geodeticDatum**, **coordinateUncertaintyInMeters**, **footprintWKT**, **coordinatePrecision**, **georeferencedBy**, **georeferenceSources**, and **georeferenceRemarks**. The underlined columns should be blank in the input dataset.
- The *omoccurduplicatelink* table from your Symbiota portal (You will need backend access to your portal to access this. Contact your portal administrator for help accessing this file).
- A simplified download of the entire Symbiota omoccurrences table containing, at minimum, the columns: **occid**, **decimalLatitude**, **decimalLongitude**, **geodeticDatum**, **coordinateUncertaintyInMeters**, **footprintWKT**, **coordinatePrecision**, **georeferencedBy**, **georeferenceSources**, and **georeferenceRemarks**.

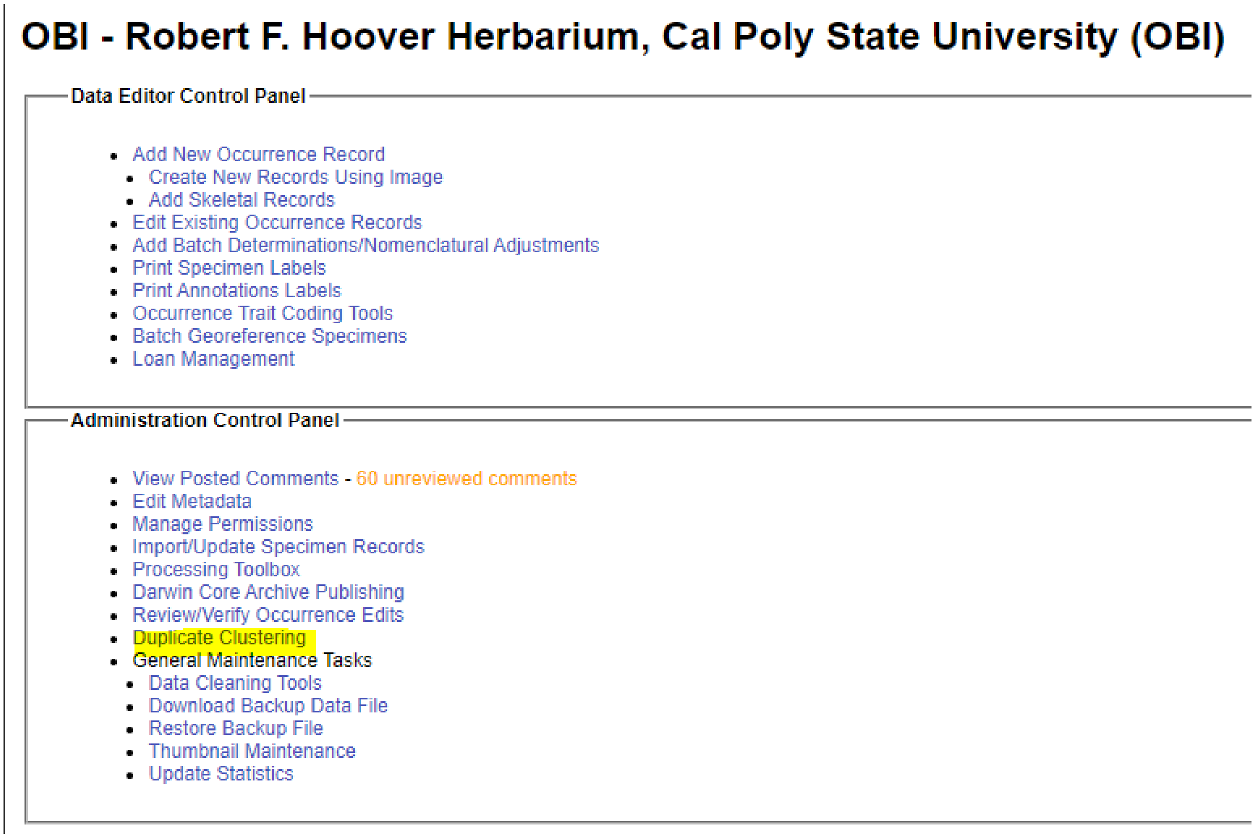
OUTPUT:

- A dataset containing georeference data (see columns in simplified omoccurrences table) for all specimens that have georeferenced duplicates in the omocurrences table.
  - Note that, in the case of multiple duplicates, one record in the target dataset may result in several records/rows in the output dataset. For this reason, cleaning steps 3-5 are provided.

These materials were made possible by National Science Foundation Award 1802312. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

STEP 1: ENSURE DUPLICATES ARE LINKED IN YOUR SYMBIOTA PORTAL

Run the duplicate clustering tool for your target collection to ensure that duplicate records are identified and stored in the *omoccurduplinks* table.



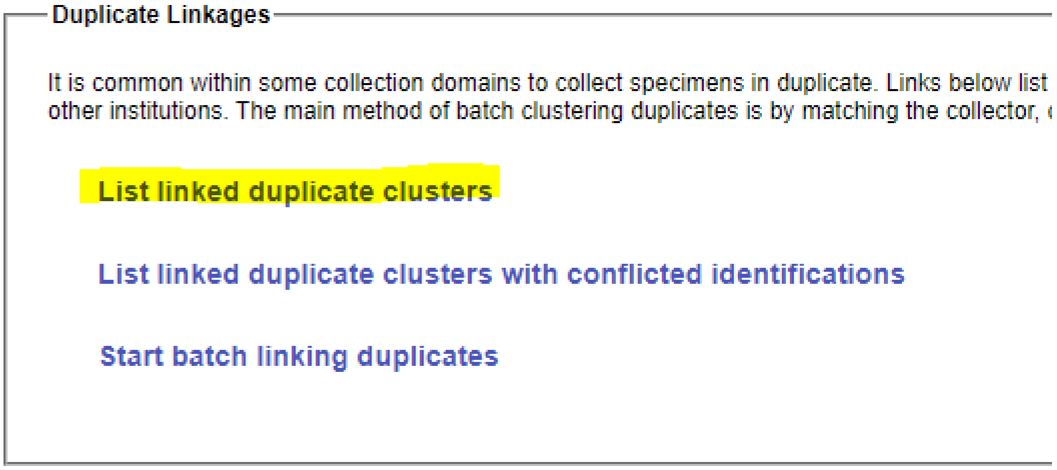
OBI - Robert F. Hoover Herbarium, Cal Poly State University (OBI)

Data Editor Control Panel

- Add New Occurrence Record
  - Create New Records Using Image
  - Add Skeletal Records
- Edit Existing Occurrence Records
- Add Batch Determinations/Nomenclatural Adjustments
- Print Specimen Labels
- Print Annotations Labels
- Occurrence Trait Coding Tools
- Batch Georeference Specimens
- Loan Management

Administration Control Panel

- View Posted Comments - 60 unreviewed comments
- Edit Metadata
- Manage Permissions
- Import/Update Specimen Records
- Processing Toolbox
- Darwin Core Archive Publishing
- Review/Verify Occurrence Edits
- Duplicate Clustering
- General Maintenance Tasks
  - Data Cleaning Tools
  - Download Backup Data File
  - Restore Backup File
  - Thumbnail Maintenance
  - Update Statistics



Duplicate Linkages

It is common within some collection domains to collect specimens in duplicate. Links below list
other institutions. The main method of batch clustering duplicates is by matching the collector,

List linked duplicate clusters

List linked duplicate clusters with conflicted identifications

Start batch linking duplicates

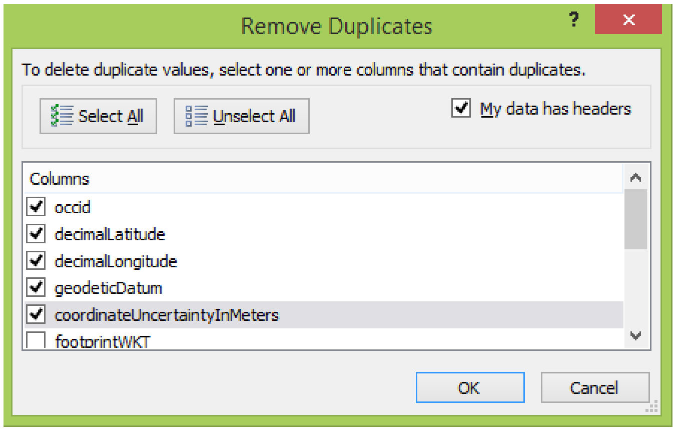
These materials were made possible by National Science Foundation Award 1802312. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Step 2: Run the code

Run the ExtractDuplicateGeoreferences R code, making sure to update the path to each input file as described in the comments of the code. The dataset of the collection you are seeking to update should be loaded as the "mycoll" data frame.

Step 3: Remove duplicate duplicates

Remove duplicate lat/longs+uncertainty in Excel:



Remove Duplicates

To delete duplicate values, select one or more columns that contain duplicates.

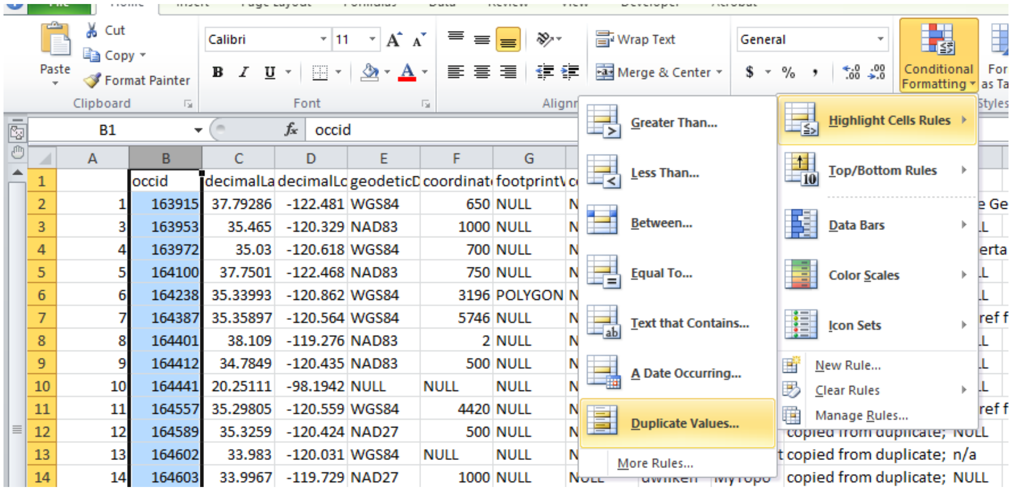
Select All | Unselect All | ☑ My data has headers

Columns
- ☑ occid
- ☑ decimalLatitude
- ☑ decimalLongitude
- ☑ geodeticDatum
- ☑ coordinateUncertaintyInMeters
- □ footprintWKT

OK | Cancel

Step 4: View duplicates

Visualize duplicates that still exist:



B1 = occid

	A	B	C	D	E	F	G
1		occid	decimalLa	decimalLc	geodeticD	coordinat	footprintV
2	1	163915	37.79286	-122.481	WGS84	650	NULL
3	3	163953	35.465	-120.329	NAD83	1000	NULL
4	4	163972	35.03	-120.618	WGS84	700	NULL
5	5	164100	37.7501	-122.468	NAD83	750	NULL
6	6	164238	35.33993	-120.862	WGS84	3196	POLYGON
7	7	164387	35.35897	-120.564	WGS84	5746	NULL
8	8	164401	38.109	-119.276	NAD83	2	NULL
9	9	164412	34.7849	-120.435	NAD83	500	NULL
10	10	164441	20.25111	-98.1942	NULL	NULL	NULL
11	11	164557	35.29805	-120.559	WGS84	4420	NULL
12	12	164589	35.3259	-120.424	NAD27	500	NULL
13	13	164602	33.983	-120.031	WGS84	NULL	NULL
14	14	164603	33.9967	-119.729	NAD27	1000	NULL

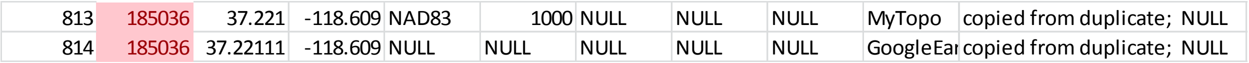
Step 5: Select best duplicate georeferences

Decide which set of duplicates to keep by following these rules. You should prioritize georeferences with:

1. purported coordinates from label
2. non-NULL georeference remarks, datum, coordinate uncertainty, etc. fields
3. named georeferencers (not just UCBerkeley, etc.)
4. smaller error radii (unless they are suspicious)

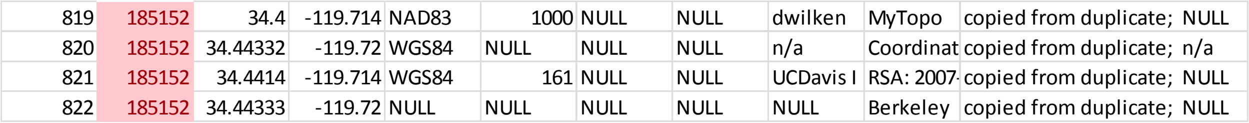
These materials were made possible by National Science Foundation Award 1802312. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Examples



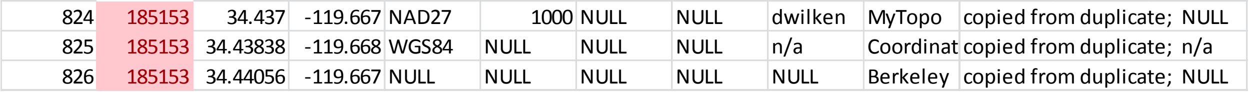
813	185036	37.221	-118.609	NAD83	1000	NULL	NULL	NULL	MyTopo	copied from duplicate; NULL
814	185036	37.22111	-118.609	NULL	NULL	NULL	NULL	NULL	GoogleEa	copied from duplicate; NULL

Keep 813 because it has datum and error radius (1000)



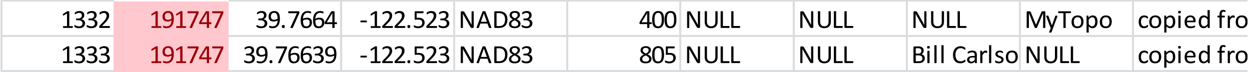
819	185152	34.4	-119.714	NAD83	1000	NULL	NULL	dwilken	MyTopo	copied from duplicate; NULL
820	185152	34.44332	-119.72	WGS84	NULL	NULL	NULL	n/a	Coordinat	copied from duplicate; n/a
821	185152	34.4414	-119.714	WGS84	161	NULL	NULL	UCDavis I	RSA: 2007-	copied from duplicate; NULL
822	185152	34.44333	-119.72	NULL	NULL	NULL	NULL	NULL	Berkeley	copied from duplicate; NULL

Keep 821 because it has the smallest error radius



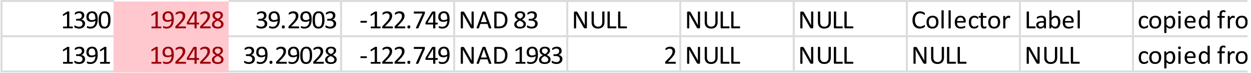
824	185153	34.437	-119.667	NAD27	1000	NULL	NULL	dwilken	MyTopo	copied from duplicate; NULL
825	185153	34.43838	-119.668	WGS84	NULL	NULL	NULL	n/a	Coordinat	copied from duplicate; n/a
826	185153	34.44056	-119.667	NULL	NULL	NULL	NULL	NULL	Berkeley	copied from duplicate; NULL

Keep 824 because it has a georeferencer (dwilken)



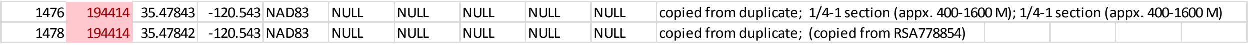
1332	191747	39.7664	-122.523	NAD83	400	NULL	NULL	NULL	MyTopo	copied fro
1333	191747	39.76639	-122.523	NAD83	805	NULL	NULL	Bill Carlso	NULL	copied fro

Keep 1333 because it has a georeferencer (Bill Carlson)



1390	192428	39.2903	-122.749	NAD 83	NULL	NULL	NULL	Collector	Label	copied fro
1391	192428	39.29028	-122.749	NAD 1983	2	NULL	NULL	NULL	NULL	copied fro

Keep 1390 because the source is the collector/specimen label

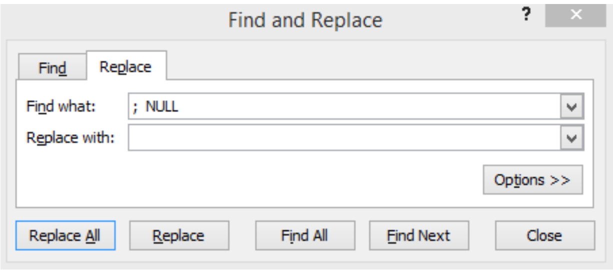


1476	194414	35.47843	-120.543	NAD83	NULL	NULL	NULL	NULL	NULL	copied from duplicate; 1/4-1 section (appx. 400-1600 M); 1/4-1 section (appx. 400-1600 M)
1478	194414	35.47842	-120.543	NAD83	NULL	NULL	NULL	NULL	NULL	copied from duplicate; (copied from RSA778854)

Keep 1476 because georeference remarks are more specific

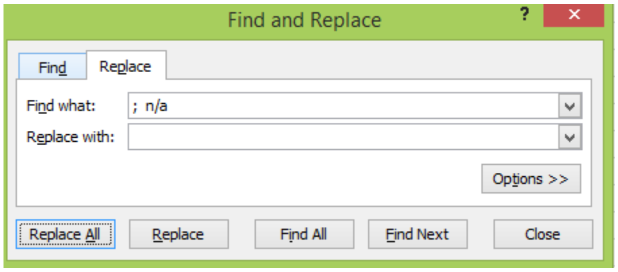
Step 6: Final cleaning

Remove "; NULL" and "; n/a" from the end of georeference remarks field (if applicable)



Find what: ; NULL

Replace with:



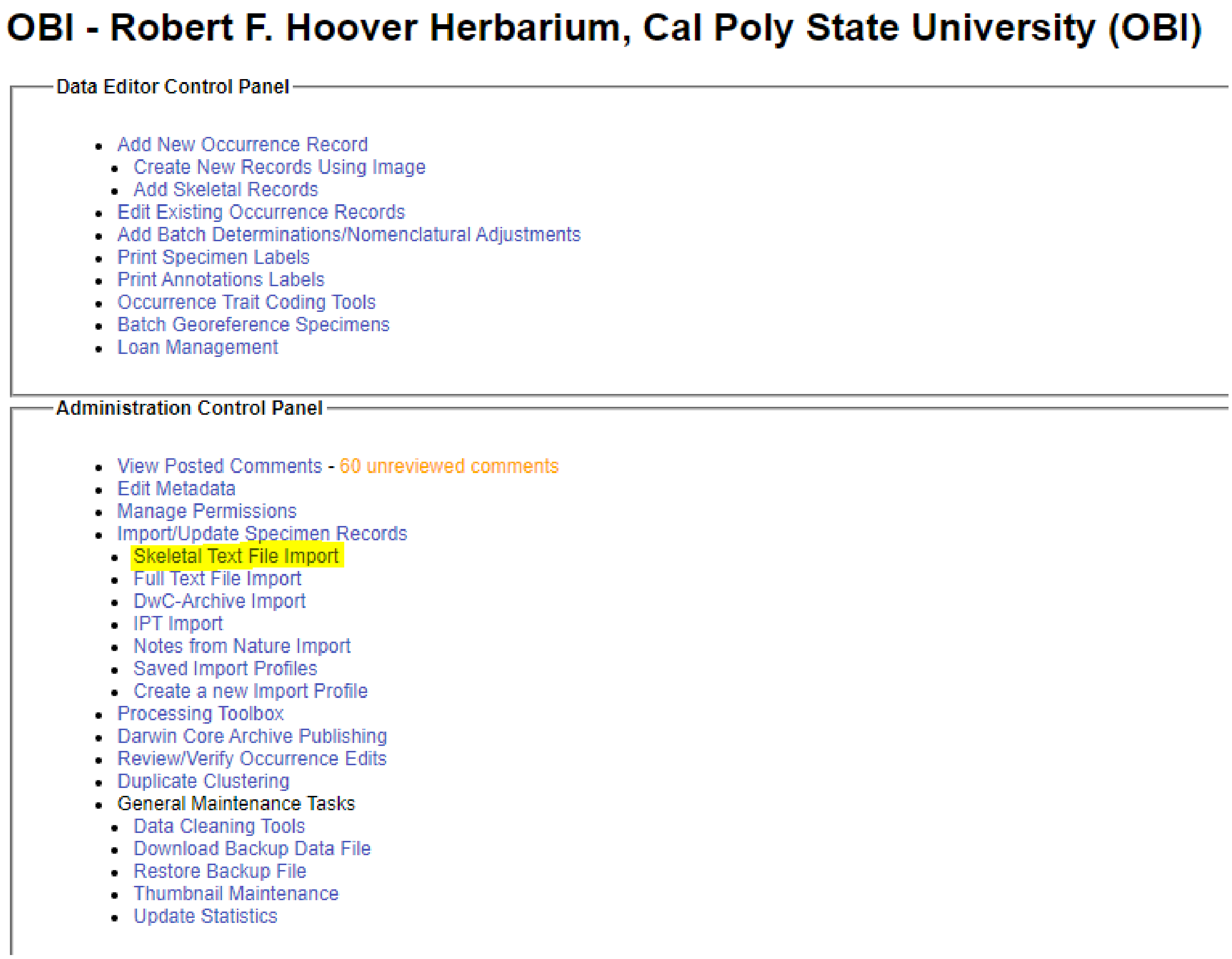
Find what: ; n/a

Replace with:

These materials were made possible by National Science Foundation Award 1802312. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Step 7: Re-integrate data into Symbiota database

Upload the data into your database using a Skeletal Text File Import tool. This will ensure that only the data from the fields in your output data file are changed and that any pre-existing data (e.g., georeferences that were added during that time it took you to run the code) are not overwritten.



OBI - Robert F. Hoover Herbarium, Cal Poly State University (OBI)

Data Editor Control Panel

- Add New Occurrence Record
  - Create New Records Using Image
  - Add Skeletal Records
- Edit Existing Occurrence Records
- Add Batch Determinations/Nomenclatural Adjustments
- Print Specimen Labels
- Print Annotations Labels
- Occurrence Trait Coding Tools
- Batch Georeference Specimens
- Loan Management

Administration Control Panel

- View Posted Comments - 60 unreviewed comments
- Edit Metadata
- Manage Permissions
- Import/Update Specimen Records
  - Skeletal Text File Import
  - Full Text File Import
  - DwC-Archive Import
  - IPT Import
  - Notes from Nature Import
  - Saved Import Profiles
  - Create a new Import Profile
- Processing Toolbox
- Darwin Core Archive Publishing
- Review/Verify Occurrence Edits
- Duplicate Clustering
- General Maintenance Tasks
  - Data Cleaning Tools
  - Download Backup Data File
  - Restore Backup File
  - Thumbnail Maintenance
  - Update Statistics

These materials were made possible by National Science Foundation Award 1802312. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.