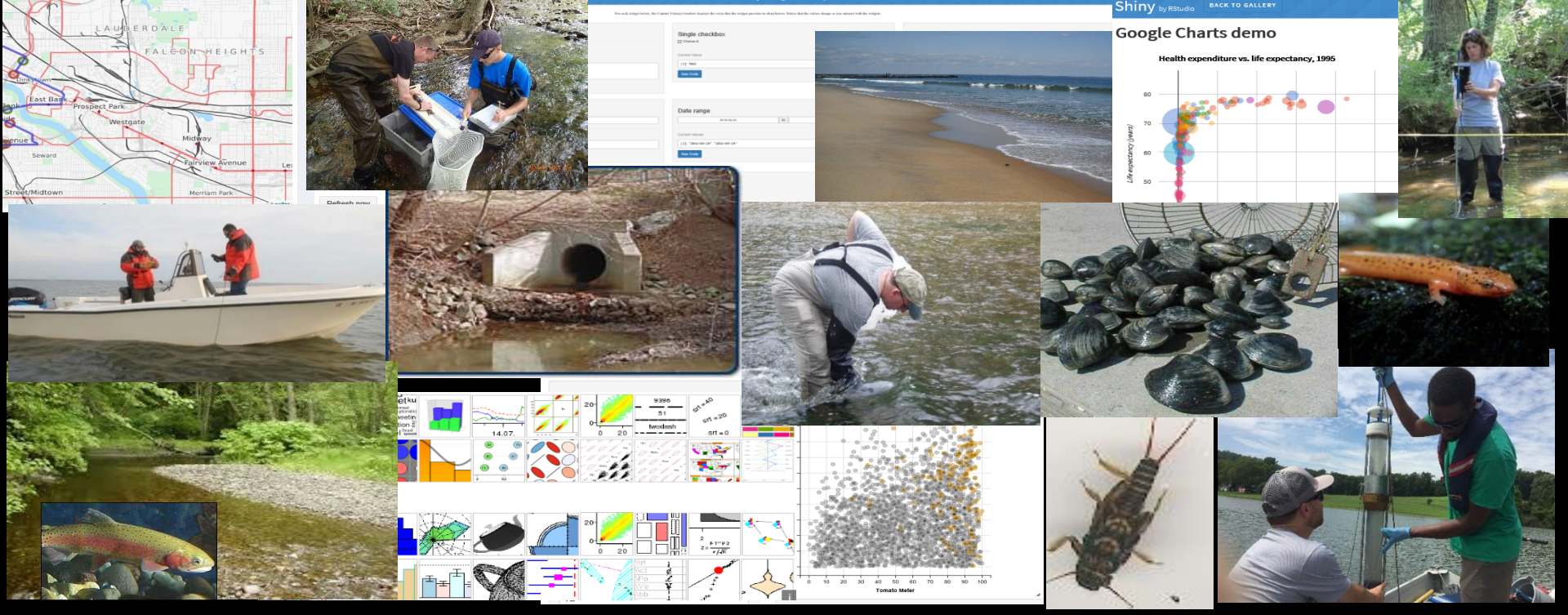


AUTOMATED ASSESSMENT: MANAGING LARGE DATA SETS WITH R

Jack Pflaumer , Biswarup Guha, Dave Rosenman

Division of Water Monitoring & Standards
New Jersey Department of Environmental Protection (NJDEP)

May 29, 2018

Overview

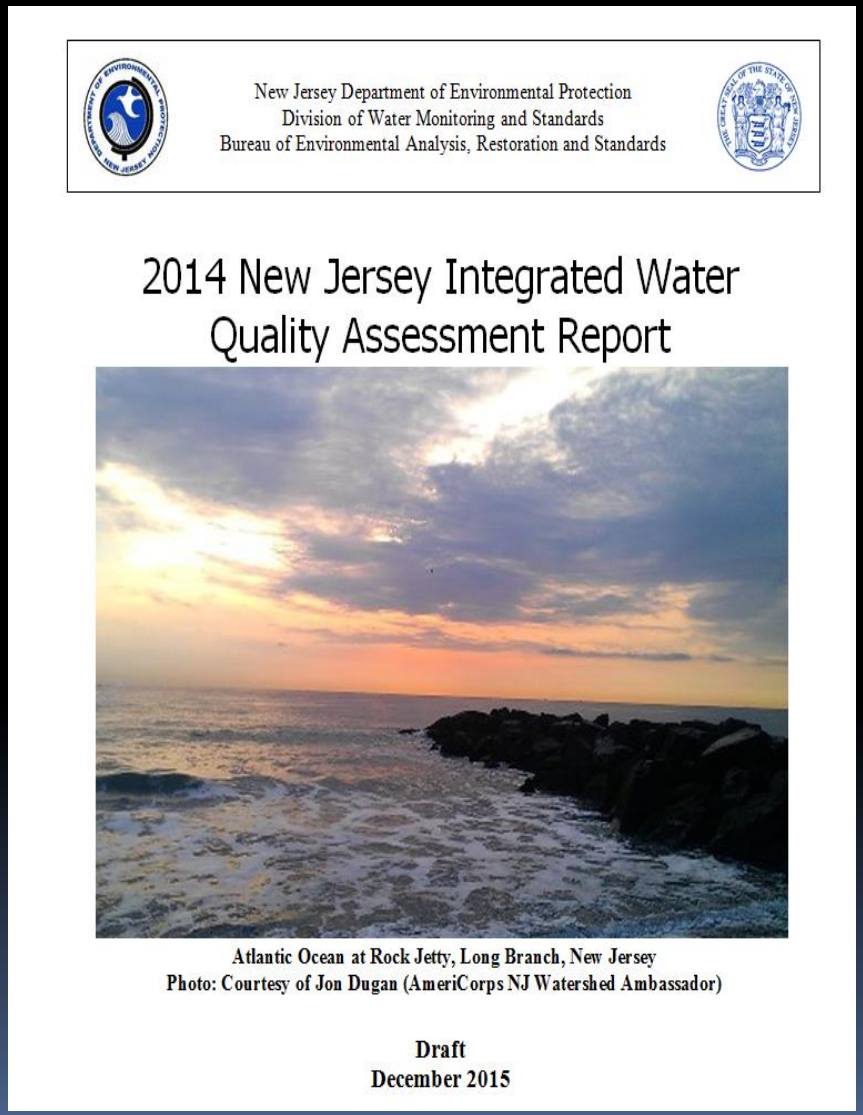
- Why Automation Needed?
- Manage Data
  - Data for Integrated Report
  - National Water Quality Portal (WQP)
  - Continuous Data Sets
- Why Use R?
  - New Jersey's Evolution to R
  - Benefits of R
  - Challenges and Lessons Learned

Why Automation Needed - Data & Assessment

- Difficult to obtain data from multiple sources and formats
- Issues with managing large data sets
- How to QA large number of stations / parameters
- Need to eliminate many repetitive data analyses steps
- Focus on water quality evaluations by scientific staff
- Improve accuracy, more complete assessments
- Improve transparency and reporting tools

Integrated Report Process

- Chemistry Data Crunching
  - 958 assessment units (HUC14)
  - 5 - >10 years of data
  - > 10,000 discrete stations
  - > 3.2 million discrete data
  - > 300 continuous monitoring stations
  - > 90 parameters
  - Biological data
- Assessments
  - Station level
  - HUC level chemistry and biological assessment rollup
  - Designated Use assessment

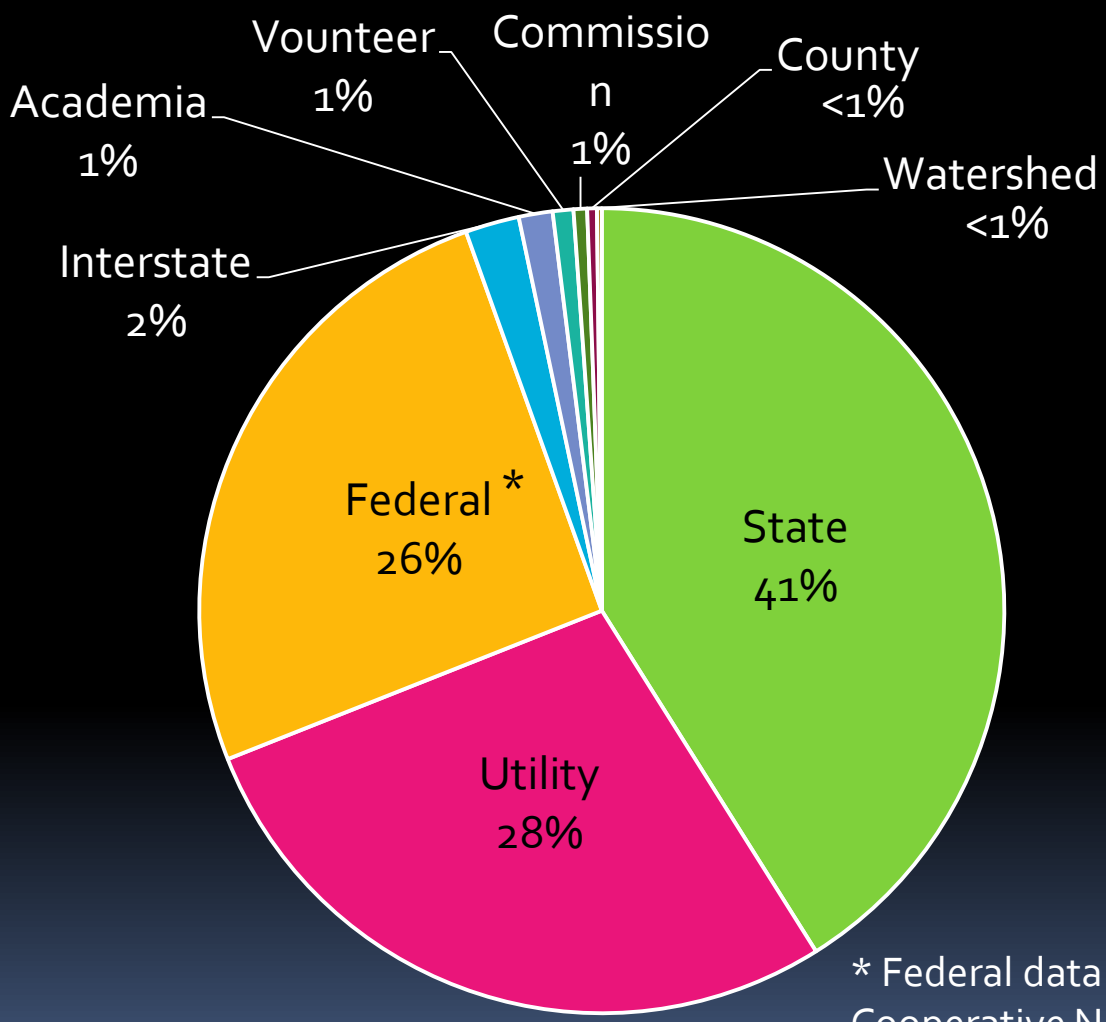

New Jersey Department of Environmental Protection
Division of Water Monitoring and Standards
Bureau of Environmental Analysis, Restoration and Standards

2014 New Jersey Integrated Water Quality Assessment Report

Atlantic Ocean at Rock Jetty, Long Branch, New Jersey
Photo: Courtesy of Jon Dugan (AmeriCorps NJ Watershed Ambassador)

Draft
December 2015

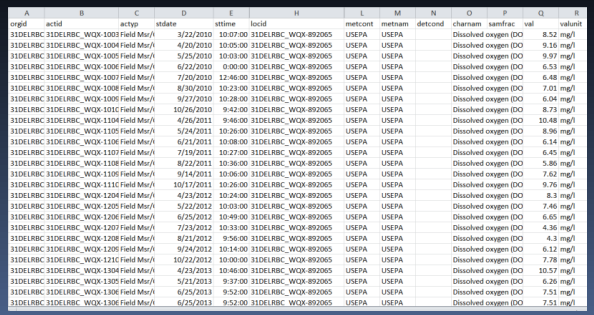
2016 NJ Integrated Report Data Sources



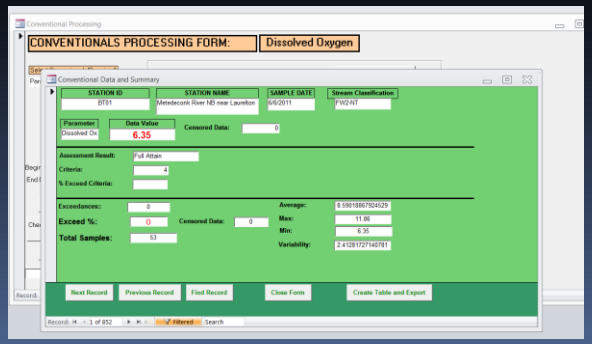
* Federal data includes:
Cooperative NJDEP/USGS
River and Stream Monitoring Network

Data Repositories

- NWIS
- STORET (pre–WQX)
  - Clearinghouse for nationwide water quality data
  - Difficult to import data
  - Difficult to retrieve data
  - Uneven participation and data availability
- Myrid of other databases (excel, access, text)



orgid	actid	actyp	stdate	sttime	locid	metcont	metnam	detcond	charnam	samfrac	val	valunit
31DELRBC	31DELRBC_WQX-1003	Field Msr/I	3/22/2010	10:07:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		8.52	mg/l
31DELRBC	31DELRBC_WQX-1004	Field Msr/I	4/20/2010	10:05:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		9.16	mg/l
31DELRBC	31DELRBC_WQX-1005	Field Msr/I	5/25/2010	10:03:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		9.97	mg/l
31DELRBC	31DELRBC_WQX-1006	Field Msr/I	6/22/2010	0:00:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.53	mg/l
31DELRBC	31DELRBC_WQX-1007	Field Msr/I	7/20/2010	12:46:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.48	mg/l
31DELRBC	31DELRBC_WQX-1008	Field Msr/I	8/30/2010	10:23:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		7.01	mg/l
31DELRBC	31DELRBC_WQX-1009	Field Msr/I	9/27/2010	10:28:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.04	mg/l
31DELRBC	31DELRBC_WQX-1010	Field Msr/I	10/26/2010	9:42:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		8.73	mg/l
31DELRBC	31DELRBC_WQX-1104	Field Msr/I	4/26/2011	9:46:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		10.48	mg/l
31DELRBC	31DELRBC_WQX-1105	Field Msr/I	5/24/2011	10:26:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		8.96	mg/l
31DELRBC	31DELRBC_WQX-1106	Field Msr/I	6/21/2011	10:08:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.14	mg/l
31DELRBC	31DELRBC_WQX-1107	Field Msr/I	7/19/2011	10:27:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.45	mg/l
31DELRBC	31DELRBC_WQX-1108	Field Msr/I	8/22/2011	10:36:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		5.86	mg/l
31DELRBC	31DELRBC_WQX-1109	Field Msr/I	9/14/2011	10:06:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		7.62	mg/l
31DELRBC	31DELRBC_WQX-1110	Field Msr/I	10/17/2011	10:26:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		9.76	mg/l
31DELRBC	31DELRBC_WQX-1204	Field Msr/I	4/23/2012	10:24:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		8.3	mg/l
31DELRBC	31DELRBC_WQX-1205	Field Msr/I	5/22/2012	10:03:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		7.46	mg/l
31DELRBC	31DELRBC_WQX-1206	Field Msr/I	6/25/2012	10:49:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.65	mg/l
31DELRBC	31DELRBC_WQX-1207	Field Msr/I	7/23/2012	10:33:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		4.36	mg/l
31DELRBC	31DELRBC_WQX-1208	Field Msr/I	8/21/2012	9:56:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		4.3	mg/l
31DELRBC	31DELRBC_WQX-1209	Field Msr/I	9/24/2012	10:14:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.12	mg/l
31DELRBC	31DELRBC_WQX-1210	Field Msr/I	10/22/2012	10:00:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		7.78	mg/l
31DELRBC	31DELRBC_WQX-1304	Field Msr/I	4/23/2013	10:46:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		10.57	mg/l
31DELRBC	31DELRBC_WQX-1305	Field Msr/I	5/21/2013	9:37:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.26	mg/l
31DELRBC	31DELRBC_WQX-1306	Field Msr/I	6/25/2013	9:52:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		7.51	mg/l
31DELRBC	31DELRBC_WQX-1306	Field Msr/I	6/25/2013	9:52:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		7.51	mg/l



The dissolved arsenic levels at these site were as follows:

01444990

2/8/2005: 0.2 ug/L
8/16/2005: 0.6 ug/L, sediment (tot rec.) 1.1 ug/g

01445100

2/16.2005: 0.3 ug/L
8/11/2005: 0.8 ug/L, sediment (tot rec.) 1.4 ug/g

01445700

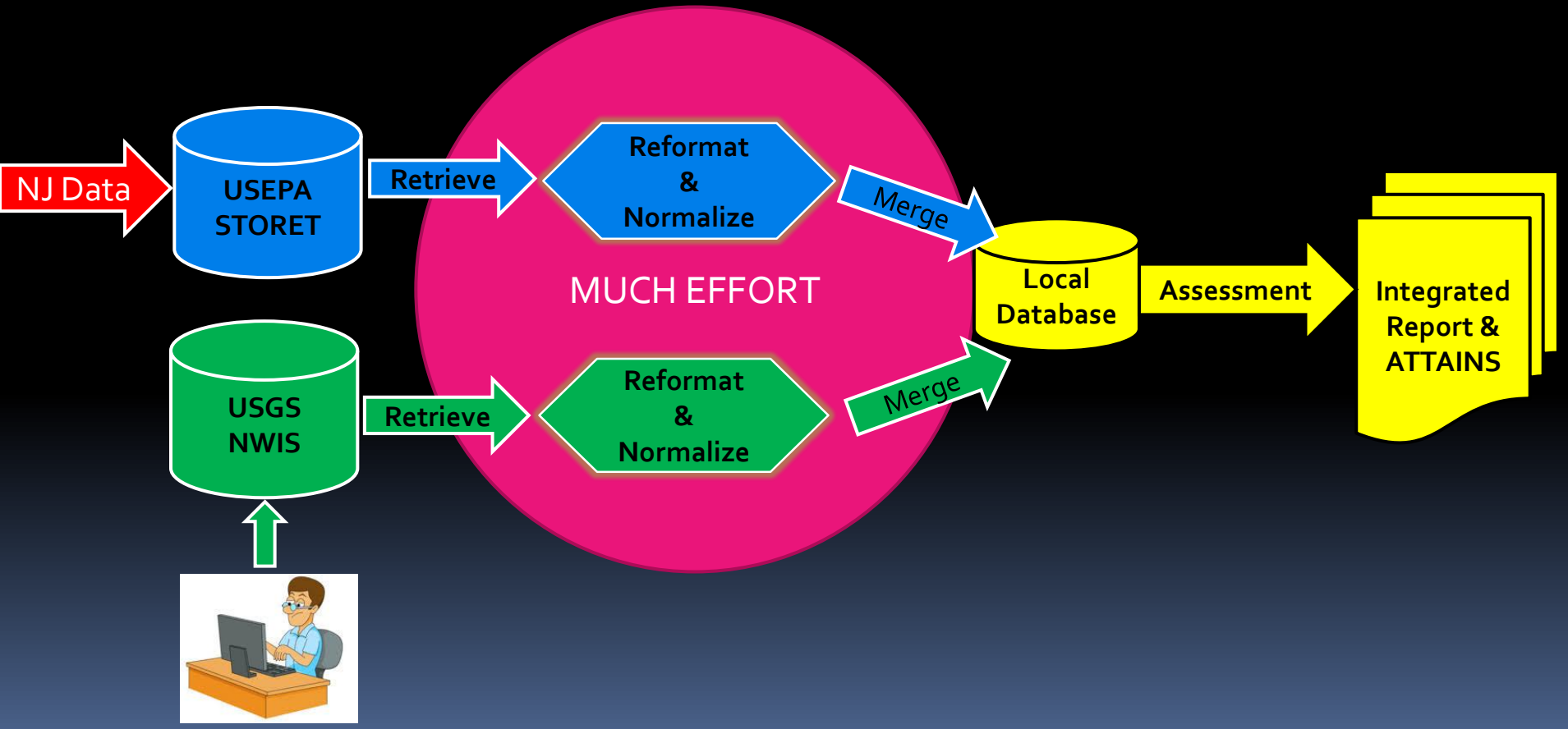
2/17/2005: 0.3 ug/L
8/24/2005: 0.7 ug/L

All total recoverable samples were <2.0 ug/L.

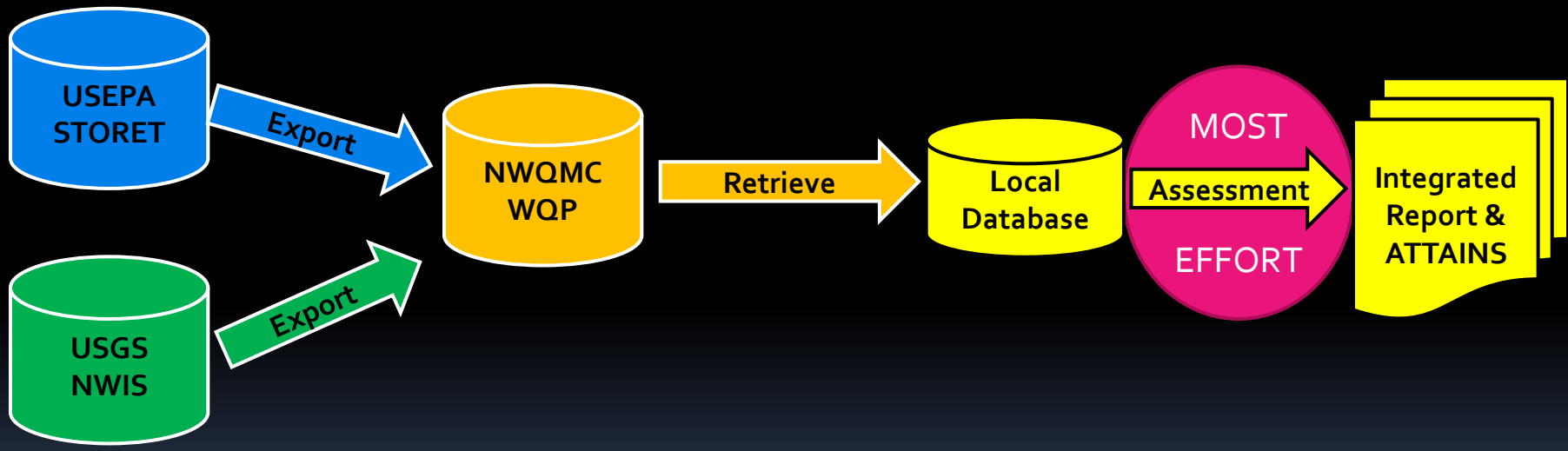
Water Quality Portal Benefits

- Since 2012: EPA (STORET), USGS (NWIS), USDA (STEWARDS)
- One Stop Shop for data (except for continuous)
- Standardized data protocol
- Easy to use interface with help desk
- Closer collaboration with stakeholders with WQX
- Publicly available and easily accessible
- NJ requires approved QAPP

Data Access & Assessment BEFORE the Water Quality Portal

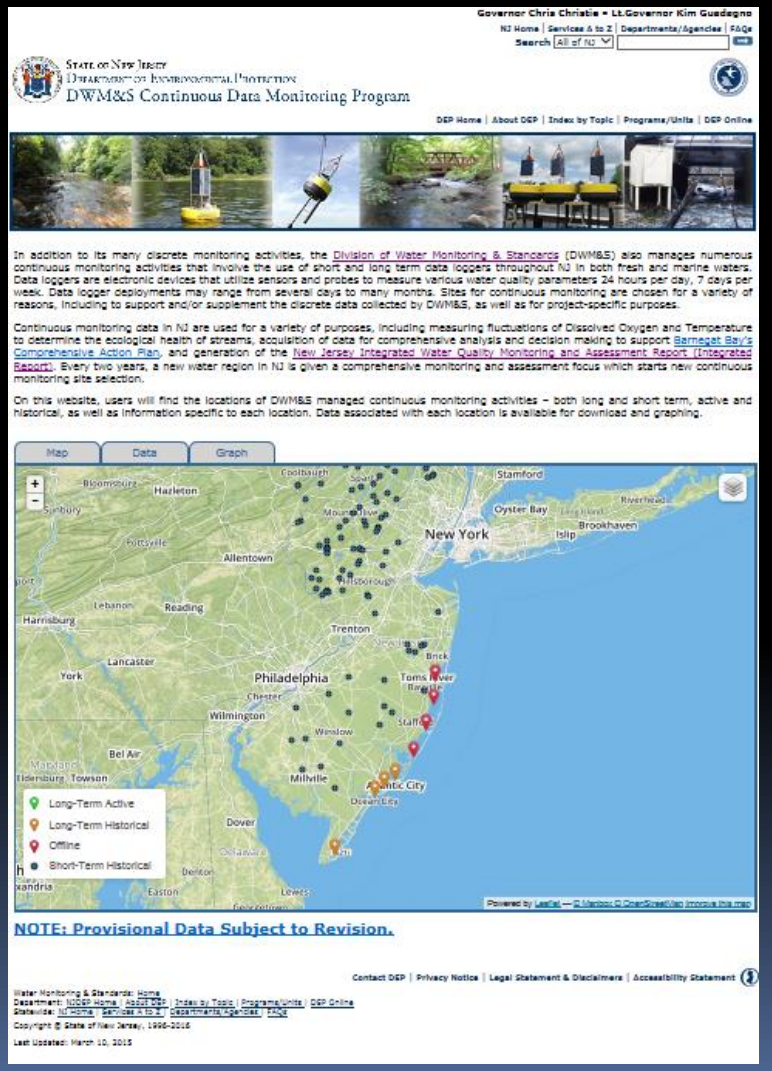


Data Access & Assessment in AFTER the Water Quality Portal

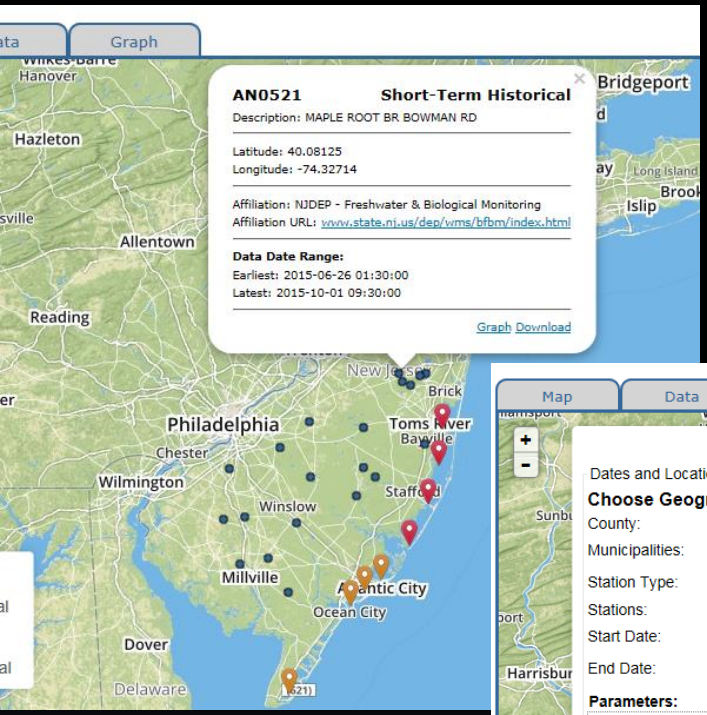


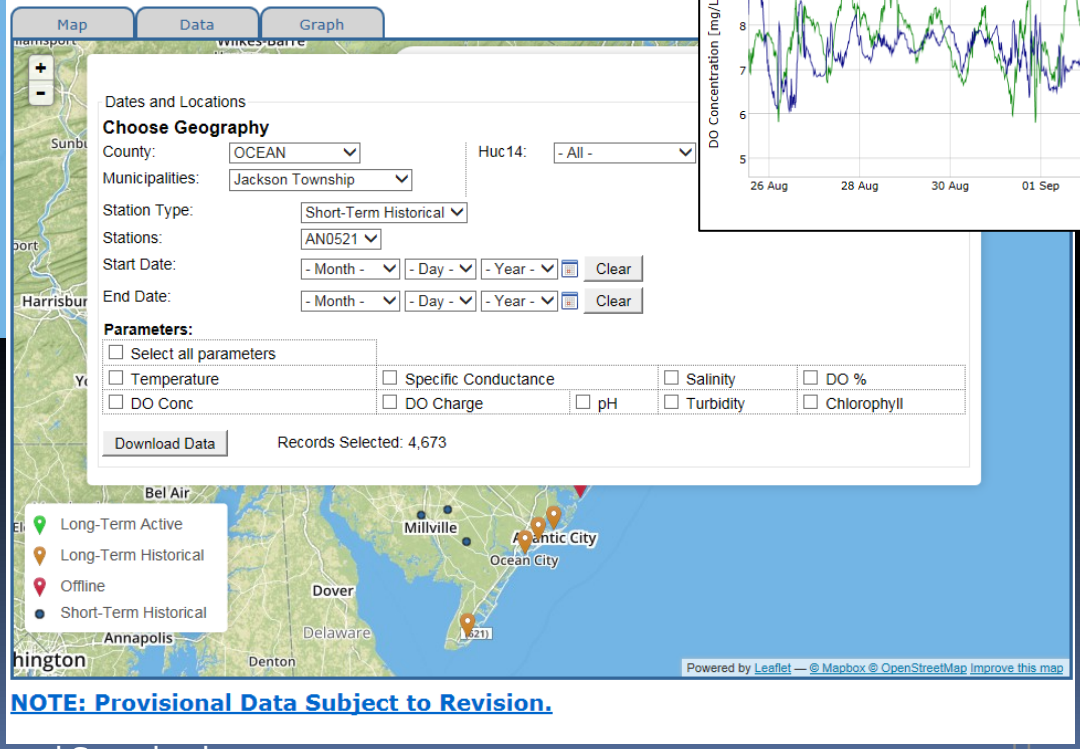
Continuous Data Solution

- National Issues with Continuous Monitoring Data
  - STORET not designed for this
  - No data standard
  - Difficulty sharing data
- NJ Solution – web accessible database (Rutgers U)
  - Interactive
  - Easy to use
  - Expandability
    - Other data generators
    - EPA is working with Rutgers

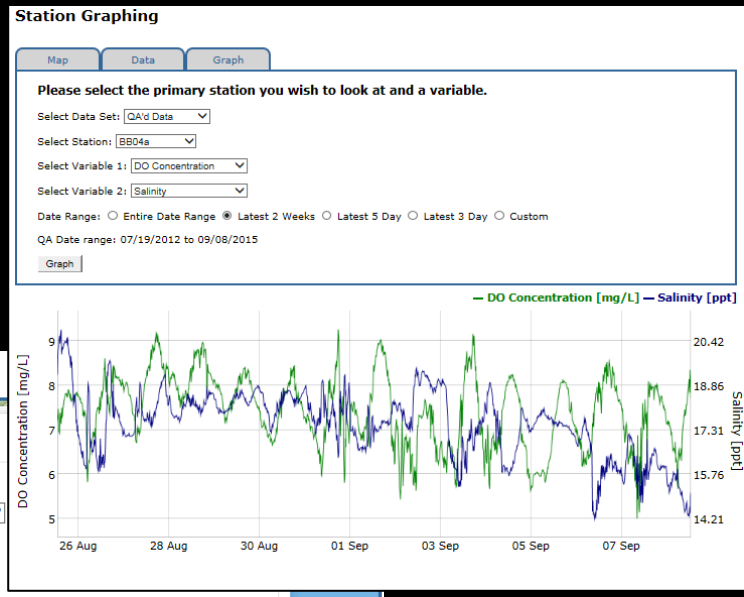


Continuous Data Solution





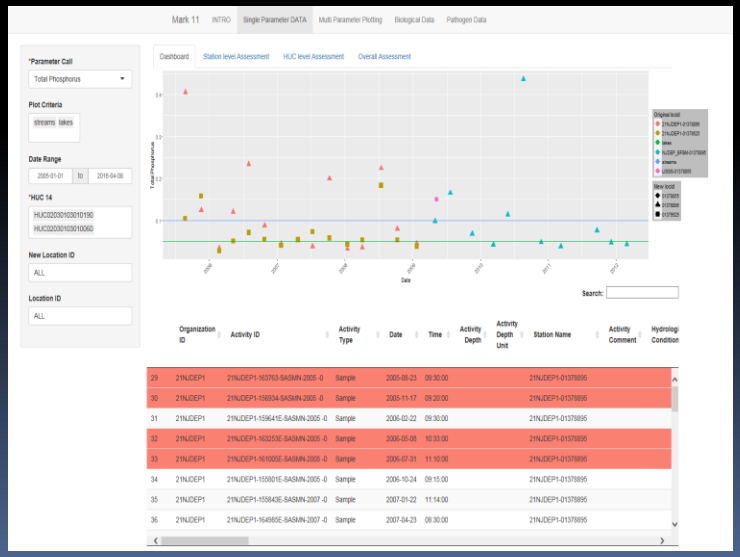
NOTE: Provisional Data Subject to Revision.

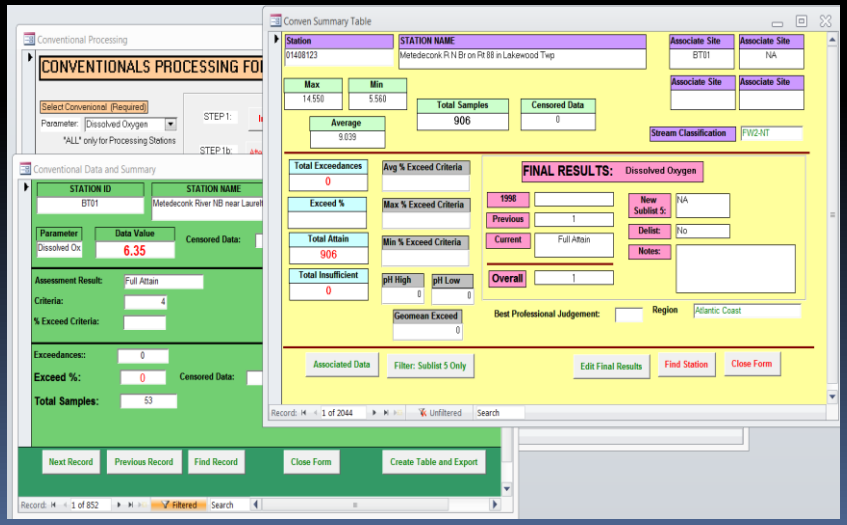


History of Automated Assessment Process

- Incremental improvements to automation process
- Excel (pre-2006)
- Access (2006)
- R (2016)

A	B	C	D	E	H	L	M	N	O	P	Q	R
orgid	actid	actyp	stdate	sttime	locid	metcont	metnam	detcond	charnam	samfrac	val	valunit
31DELRBC	31DELRBC_WQX-1003	Field Msr/I	3/22/2010	10:07:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		8.52	mg/l
31DELRBC	31DELRBC_WQX-1004	Field Msr/I	4/20/2010	10:05:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		9.16	mg/l
31DELRBC	31DELRBC_WQX-1005	Field Msr/I	5/25/2010	10:03:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		9.97	mg/l
31DELRBC	31DELRBC_WQX-1006	Field Msr/I	6/22/2010	0:00:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.53	mg/l
31DELRBC	31DELRBC_WQX-1007	Field Msr/I	7/20/2010	12:46:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.48	mg/l
31DELRBC	31DELRBC_WQX-1008	Field Msr/I	8/30/2010	10:23:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		7.01	mg/l
31DELRBC	31DELRBC_WQX-1009	Field Msr/I	9/27/2010	10:28:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.04	mg/l
31DELRBC	31DELRBC_WQX-1010	Field Msr/I	10/26/2010	9:42:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		8.73	mg/l
31DELRBC	31DELRBC_WQX-1104	Field Msr/I	4/26/2011	9:46:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		10.48	mg/l
31DELRBC	31DELRBC_WQX-1105	Field Msr/I	5/24/2011	10:26:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		8.96	mg/l
31DELRBC	31DELRBC_WQX-1106	Field Msr/I	6/21/2011	10:08:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.14	mg/l
31DELRBC	31DELRBC_WQX-1107	Field Msr/I	7/19/2011	10:27:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.45	mg/l
31DELRBC	31DELRBC_WQX-1108	Field Msr/I	8/22/2011	10:36:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		5.86	mg/l
31DELRBC	31DELRBC_WQX-1109	Field Msr/I	9/14/2011	10:06:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		7.62	mg/l
31DELRBC	31DELRBC_WQX-1110	Field Msr/I	10/17/2011	10:26:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		9.76	mg/l
31DELRBC	31DELRBC_WQX-1204	Field Msr/I	4/23/2012	10:24:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		8.3	mg/l
31DELRBC	31DELRBC_WQX-1205	Field Msr/I	5/22/2012	10:03:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		7.46	mg/l
31DELRBC	31DELRBC_WQX-1206	Field Msr/I	6/25/2012	10:49:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.65	mg/l
31DELRBC	31DELRBC_WQX-1207	Field Msr/I	7/23/2012	10:33:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		4.36	mg/l
31DELRBC	31DELRBC_WQX-1208	Field Msr/I	8/21/2012	9:56:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		4.3	mg/l
31DELRBC	31DELRBC_WQX-1209	Field Msr/I	9/24/2012	10:14:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.12	mg/l
31DELRBC	31DELRBC_WQX-1210	Field Msr/I	10/22/2012	10:00:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		7.78	mg/l
31DELRBC	31DELRBC_WQX-1304	Field Msr/I	4/23/2013	10:46:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		10.57	mg/l
31DELRBC	31DELRBC_WQX-1305	Field Msr/I	5/21/2013	9:37:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		6.26	mg/l
31DELRBC	31DELRBC_WQX-1306	Field Msr/I	6/25/2013	9:52:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		7.51	mg/l
31DELRBC	31DELRBC_WQX-1306	Field Msr/I	6/25/2013	9:52:00	31DELRBC_WQX-892065	USEPA	USEPA		Dissolved oxygen (DO		7.51	mg/l

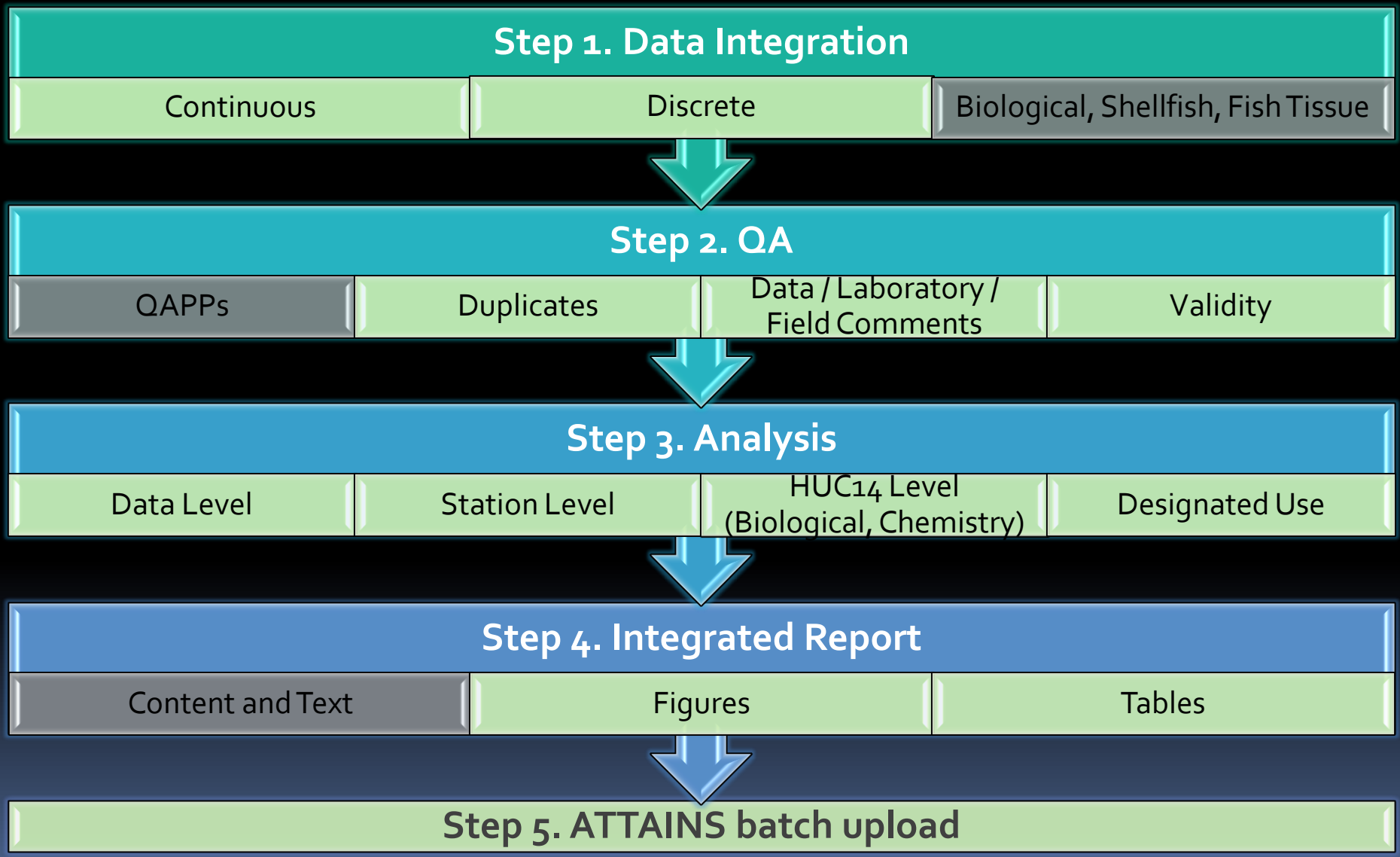




Why Manage Data with R

- Free software
- Easily handles large datasets, easy to use
- No special data management system needed
- Store data in ordinary ascii or binary files
  - Large datasets – binary format improved efficiency
- Automated downloads (WQP, USGS)
- Vast community and resources available
- Can do everything with R: data management, assessment, reporting
- Integrates with other languages and platforms

NJ IR Automated Process

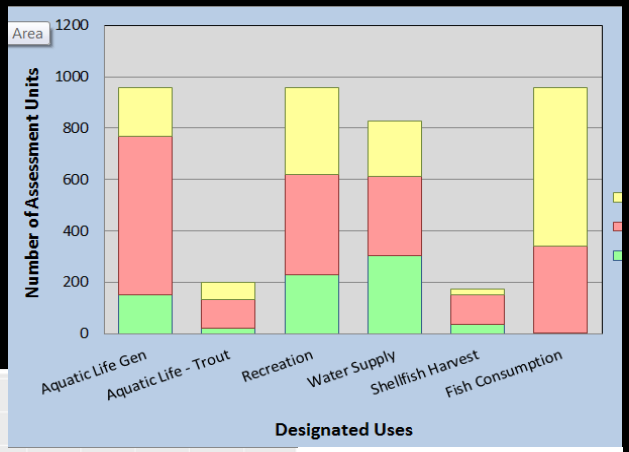


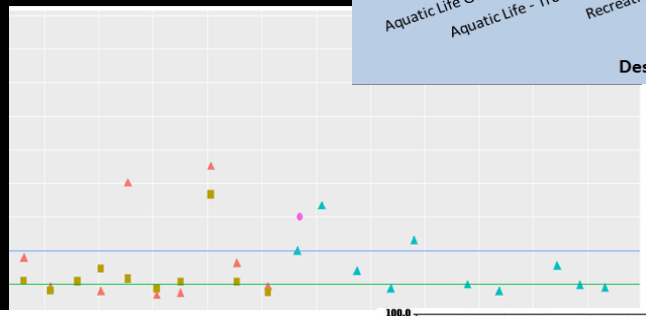
Automated

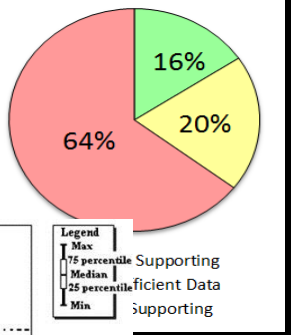
Manual

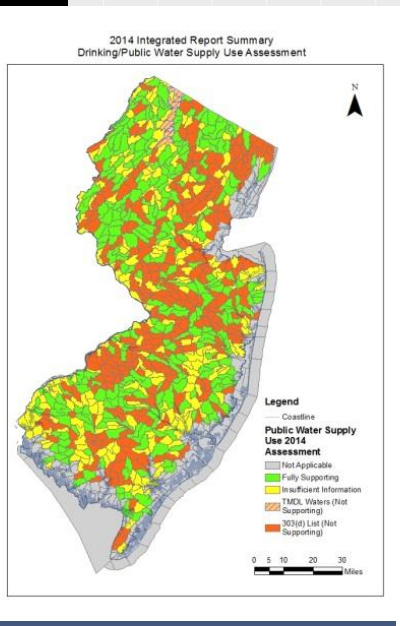
Reporting Tools

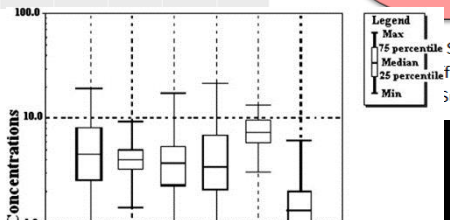
- Maps
- Tables
- Piecharts,
- Bar Charts
- Scatter Plots
- Box Plots
- GIS
- Etc….











2014 New Jersey Integrated Report — Appendix B: 2014 Final 303(d) List of Water Quality Limited Waters with Sublist 5 Subpart and Priority Ranking for TMDL Development — NJ Department of Environmental Protection May 2017

WMA	Assessment Unit Number	Assessment Unit Name	Parameter	Station Number	Cycle 1st Listed	Designated Use	Sublist 5 Subpart (A, R, L)	Priority Ranking for TMDL
15	02040302020030-01	Absecon Creek (AC Reservoirs) (gage to SB)	Mercury in Fish Tissue	Atlantic City Reservoir (Upper)	2014	Fish Consumption		Low
15	02040302020030-01	Absecon Creek (AC Reservoirs) (gage to SB)	Mercury in Water Column	01410455	2010	Water Supply		Low
15	02040302020040-01	Absecon Creek (below gage)	Mercury in Fish Tissue	Atlantic City Reservoir (Lower)	2008	Fish Consumption		Low
15	02040302020040-01	Absecon Creek (below gage)	Oxygen, Dissolved	833_2401	2004	Aquatic Life		Medium
15	02040302020040-01	Absecon Creek (below gage)	Total Coliform	Shellfish Network	2014	Shellfish		Medium
15	02040302020010-01	Absecon Creek NB	Mercury in Fish Tissue	Atlantic City Reservoir (Lower)	2008	Fish Consumption		Low
15	02040302020010-01	Absecon Creek NB	pH	ANOABGSP	2006	Aquatic Life		Medium
15	02040302020020-01	Absecon Creek SB	Mercury in Water Column	01410455	2008	Water Supply		Low
14	02040301160110-01	Albertson Brook / Gun Branch	pH	0140941020, 0140940970, NALDEREL	2006	Aquatic Life		Medium
11	02040105210010-01	Alexauken Ck (above 74d 55m)	pH	01461900	2014	Aquatic Life		Medium
11	02040105210010-01	Alexauken Ck (above 74d 55m)	Temperature, water	01461900	2006	Aquatic Life - Trout	R	Medium
11	02040105210020-01	Alexauken Ck (below 74d 55m to 11BA06)	Arsenic	01461840	2012	Water Supply		Low
11	02040105210020-01	Alexauken Ck (below 74d 55m to 11BA06)	Escherichia coli	01461840	2012	Recreation	R	Medium
11	02040105210020-01	Alexauken Ck (below 74d 55m to 11BA06)	pH	01461900	2012	Aquatic Life, Aquatic Life - Trout		Medium
11	02040105210020-01	Alexauken Ck (below 74d 55m to 11BA06)	Temperature, water	01461900	2006	Aquatic Life - Trout	R	Medium
17	02040206060020-01	Alloway Ck (above Alloway-Woodstown Rd)	Arsenic	01482880	2008	Water Supply	A	Low
17	02040206060020-01	Alloway Ck (above Alloway-Woodstown Rd)	Total Suspended Solids (TSS)	01482880	2008	Aquatic Life		Medium
17	02040206060090-01	Alloway Ck (below HancocksBr) to Salem R	PCB in Fish Tissue	Delaware River Tribs to Head of Tide	2006	Fish Consumption	L	Low
17	02040206060090-01	Alloway Ck (below HancocksBr) to Salem R	Total Coliform	Shellfish Network	2014	Shellfish		Medium
17	02040206060080-01	Alloway Ck (HancocksBridge to NewBridge)	PCB in Fish Tissue	Delaware River Tribs to Head of Tide	2006	Fish Consumption	L	Low
17	02040206060080-01	Alloway Ck (HancocksBridge to NewBridge)	Total Coliform	Shellfish Network	2014	Shellfish		Medium

Page 1 of 101

R Automation Benefits

- Improved Efficiency
  - ~ 70% time savings for NJ
- More Time for Comprehensive Assessments
- Improved Accuracy
- Improved Validation
- More Information for Technical Decisions
- User Friendly Visualization Tools
- Overall Improved Transparency of Process

Challenges with automation:
- Programming Skills & Time
- Validation & QA : requires scientific staff verification

Lessons Learned

- Use the resources on the internet
- New features / packages / tools make a difference
- Train more than one person, get a computer programmer (non-environmentalist) or contract out
- It takes much longer time to program if doing it in-house
- Not a one-time deal: continuous updates needed
- IT Department approvals can be a long process
- Lookup tables – accuracy is critical
- Inconsistent data standards – continued improvement required

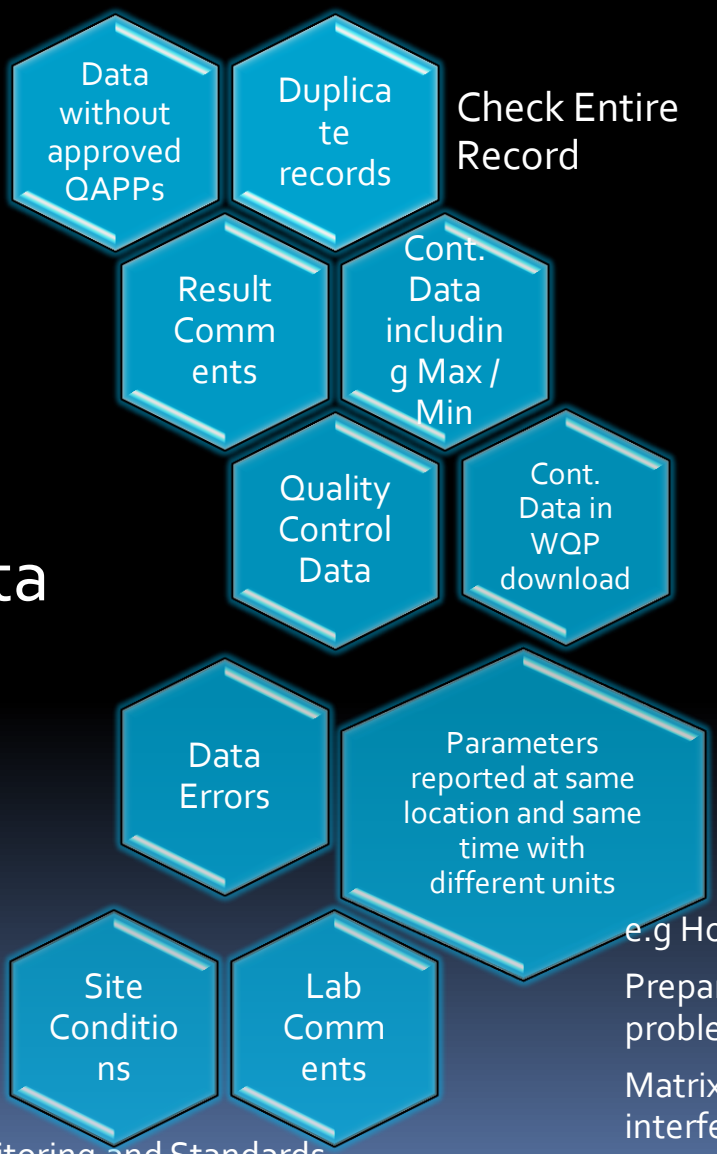
Questions

Point of Contacts:
Jack Pflaumer: (609) 633-0499 jack.pflaumer@dep.nj.gov
Roop Guha: (609) 292-1592 biswarup.guha@dep.nj.gov

New Jersey Department of Environmental Protection (NJDEP)
Division of Water Monitoring and Standards (WM&S)

Quality Assurance Tools

QA Tools Remove Data



Check Entire Record

e.g Holding time

Preparation problems

Matrix interferences

QA Tools Substitute Record

- Substitute Censored Data
- Normalize reporting unit and parameter names
- Flag preliminary and estimated Data