Data Flow in the EMAP Western Pilot Study

September 10, 1999

1. Introduction

Data collected by the EMAP Western Pilot Study (EMAP-West) are used primarily by Study participants to fulfill Study objectives**C**assessments of the environmental conditions of the region. These assessments require an early exchange of preliminary data among all Study participants and partners. Further, because these data are: (1) collected under a consistent design and consistent methods over broad regions, (2) of high-quality, and (3) well-described, they are valuable to many researchers and managers outside the Study. After quality assurance and Study analyses are completed, the data will be moved to the public domain in a form that is useful to a wide range of users and will remain accessible for the long term.

The Data Policy Statements for the EMAP Western Pilot Study outline the expectations for data sharing and the time line. This Data Flow document provides details on the flow of data and reinforces the notion that as soon as data become available to any EMAP-West participant (ORD, Region, state, tribe), those data should be shared with all participants. The time line provides target goals for data exchange. These are sometimes difficult to meet for data coming from partners outside EPA who have different contracts and procedures. Groups that cannot make these time lines should inform the Chair of the Steering Committee and the Chair of the Technical Subcommittee.

2. Data Systems

EMAP-West data, or information about the data, are slated for three information management systems: EMAP=s analytical databases and public web site, the Office of Water=s STORET, and the Office of Research and Development (ORD) Environmental Information Management System (EIMS) that holds information about data sets. Additionally, portions of these data will be stored in various state and tribal data systems.

EMAP

EMAP data management is described in the *EMAP Information Management Plan: 1998**B** 2001* (http://www.epa.gov/emap/html/pubs/docs/imsumm.html).

STORET

One objective of the EMAP geographic studies like EMAP-West is to build the capacity of groups within the area for long-term environmental monitoring, analysis, and data management. STORET, which is used by EPA Regions and many state agencies for regional and local data, will aid in fulfilling this objective. ORD will load EMAP-West Surface Waters and Coastal data to the central STORET warehouse. The Western Ecology Division (WED) in Corvallis will load Surface Waters data and the Atlantic Ecology Division (AED) in Narragansett will load Coastal data. The value of STORET will be increased by the addition of EMAP data. The capabilities of STORET will be enhanced (e.g., handling probability-based survey data, enhanced data download capability, and the addition of FGDC federal metadata standards). This collaboration will also create opportunities for coordination on data standards (e.g., Integrated Taxonomic Information System, or ITIS, species codes). Users will benefit from having a consistent database (STORET) for water quality and biological data from many sources.

During the EMAP-West Project, ORD will operate two local copies of STORET (WED for Surface Waters data and AED for Coastal data). Following the quality-assurance and analysis phases both the WED and AED STORET data bases will be uploaded to the central STORET Warehouse. The STORET Warehouse will serve as a public access portal to these data.

STORET was designed to provide for the full documentation of quality-assured environmental water quality and biological monitoring data. STORET was not designed to support day-to-day operations nor was it designed specifically to support EMAP-West data analysis. Therefore, field data will first be housed in EMAP analytical databases at WED and AED. These databases are designed to fully support the quality assurance processes and statistical analyses employed in EMAP regional assessments and will serve as the databases of record during the five years of the EMAP-West project. These field data will be made freely available to all EMAP partners for preliminary analysis; however, EMAP partners should acknowledge the preliminary nature of these data and should not attempt to maintain these data simultaneously in other systems during the quality-assurance and analysis phases.

EPA Regions and many state agencies operate local copies of the STORET Oracle database that they use to house regional and local data. At the completion of the analysis phase, regions and states that choose to load some of their EMAP-West data into their local copies of STORET will be provided with standard codes and descriptions by ORD with the assistance of EPA Headquarters STORET. Use of these standard codes is necessary to avoid confusion and to foster data exchange.

STORET enhancements

1. Enhance capability to store indicator data and data from probabilistic surveys. WED will work with EPA HQ STORET.
2. Add fields to allow compliance with Federal Geographic Data Committee (FGDC) 1998 metadata standard
3. Develop computer routines to upload Surface Water data from WED
4. Develop computer routines to upload Coastal data from AED
5. Make central STORET warehouse data available on the web

ORD is committed to a close relationship with STORET, including creating and maintaining local STORET databases. Some of the enhancements to STORET are critical, while other enhancements could simplify EMAP-West tasks. The ORD ability to meet the STORET commitment will depend on the time frame of the required STORET enhancements. Work on these depends on the budget and other priorities; hence, the time line is uncertain. Since STORET is still in the development stage as far as EMAP data are concerned, it is hard to predict what type of resources (people) it will take to meet the STORET commitment.

STORET activities

1. EPA HQ STORET will train AED and WED personnel in the database structure and procedures of STORET. A technical workshop is planned for fall 1999, after which we will be better able to predict what resources are needed and what the timeline will be.
2. EMAP-West may need to collect additional data to fulfill the metadata requirements for STORET. We can create additional data forms if necessary.
3. EPA HQ STORET will work with EMAP to develop all the codes for the EMAP sampling processes to complete the "Preferences and Defaults" section of STORET. We will them publish them, anyone wishing to store EMAP data in their copy of STORET will at least have the proper codes, and use the proper equipment, with the proper citations.
4. EPA HQ STORET will work with EMAP to develop loading and distribution routines such that, data can be moved between STORET sources.
5. EPA HQ STORET will work with EMAP to ensure that field sheets have all codes and information necessary to allow scanning and storage of complete sample and measurement profiles.
6. After data are final, they will be loaded into the local ORD copies of STORET at WED and AED. They will then be uploaded to the central STORET warehouse for public access.

The above activities should not affect field sampling schedules. If the States that wish to have their data stored in their copy of STORET are willing to wait until all the data are final, we can provide them with the upload batch files to run into their copy of STORET.

EIMS (Environmental Information Management System)

EIMS is an ORD system that stores information about ORD (and other) data**C**what data exist, where they can be found, and so on. The EMAP Data Directory, used for keeping track of the EMAP-West data sets, is periodically uploaded to the EIMS. EIMS makes the existence of EMAP data sets more widely known. EIMS is also available to use as a directory for the myriad non-EMAP data used in EMAP-West assessments. Any EPA Region, state, or ORD lab can make EIMS directory entries for these external (non-EMAP) data sets, as Region 10 has been doing in the last few years.

3. Data Flow

Figure 1 shows the basic flow of data. The Surface Waters, Coastal, and Landscape Groups consist of ORD/Region partners and will include states and tribes in the field data collection. The resource data centers bring together all the field data and the results data from samples sent to analytical laboratories (such as benthic invertebrate samples). The data centers are: **Surface Waters**: EPA Western Ecology Division (WED) in Corvallis, OR; **Coastal**: Southern California Coastal Water Research Project (SCCWRP) in Westminster, CA; and **Landscape**: EPA Environmental Sciences Division (ESD) in Las Vegas, NV. These centers are a source of preliminary data to all study participants and partners. EMAP information management and the EMAP web site are coordinated by the EPA Atlantic Ecology Division (AED). The EMAP internal web site is used for storing preliminary data. The EMAP public web site on the EPA public web server is used to distribute data and information to researchers and managers in other organizations and to other data systems. STORET will be used as a long-term archival system for the Surface Waters and Coastal data. Data sets on the EMAP public site and in STORET must be 100% quality-assured and be accompanied by metadata. **Table 1** describes what occurs during each step of the data flow.

4. Outline of Data Flow for Surface Waters Group

(See **Figure 2**)

1. Field crews from lead state agency fill out paper forms supplied by Western Ecology Division (WED), keep a copy, and send original to WED for scanning.
2. Field crews send samples (macroinvertebrates, chemistry) to analytical labs.
3. WED scans the field paper forms, verifies accuracy; sends flat ASCII files back to the lead state agency for data verification (within 6 months of field sampling). At this point, the state agency can use these data for whatever purpose they may have. If they choose to load these data to their local copy of STORET; they have to be sure not to upload them to the central

STORET warehouse because that step will be done centrally by WED when all data files, including lab results and derived data, are complete.

4. WED puts validated field data into Surface Waters database at WED.
5. WED makes validated field data available to all partners (either by email, ftp, or moving them to a password-protected area of the EMAP web site).
6. WED starts preparing metadata files (with the help of the field crews).
7. Analytical labs send results back to whoever is the Project Officer for those samples; also sends a copy to WED. A copy could also be sent to the States so that they have the raw lab data while they are waiting for completion of the WED validation procedures.
8. Project Officer verifies that analytical lab work was done in accordance with contract requirements; notifies WED. This should be completed within 12 months of field collection (depends on contract language). Water chemistry data will probably be ready sooner than biological data.
9. WED sends validated lab results to lead state agency
10. WED incorporates lab results into Surface Waters database.
11. WED makes validated lab results data available to all partners (either by email, ftp, or moving them to a password-protected area of the EMAP web site), within 15 months of field collection.
12. WED continues metadata preparation.
13. Surface Waters analytical team (WED, Regions, states, tribes) analyzes data, using the Surface Waters database and develops indices and other derived data. If it would help data access, this Surface Waters database can be put on the EMAP internal web server with password protection so that it can be used only by the Surface Waters analytical team. Otherwise, data files can be exchanged by email.
14. WED incorporates derived data into database.
15. WED finishes metadata files.
16. WED sends all data and metadata to AED to be put on the internal EMAP web site.
17. WED makes derived data available to all partners (either by email, ftp, or moving them to a password-protected area of the EMAP web site).
18. WED loads all data and data description information to the WED copy of STORET.
19. AED moves data and metadata from internal web site to public EMAP web site (within 24 months of field collection).
20. WED uploads data from WED copy of STORET to the central STORET warehouse (within X months of field collection; depends on completion of STORET enhancements; see Sect. 2).

Notes on Figure 2, Surface Waters Data Flow

1. Field form design is done at WED. Production/distribution can continue at WED, or Forms can be output in .pdf format and printed remotely. Additional software would need to be purchased. Labels/Sample tracking materials would continue at WED.

2. Raw data represents an electronic form of what is submitted on the field forms. Verification at this point is to ensure the electronic form is the same as the paper form. This data is available to the States via the Web Site and/or STORET. This is available within 6 months of submitting field forms.
3. SAS and ASCII are two formats that are currently offered. It is not difficult to add additional format options, i.e. Excel.
4. Raw data documentation is limited to methods manuals and brief variable definitions.
5. Validated data is combined data from field forms, lab and other analyses, and metric calculations. The speed at which this data is completed is in a large part dictated by the time it takes for others (labs, etc.) to submit results. Locally maintained processes are far more predictable. Physical Habitat data, for example can be validated at the rate of ten sites per week.
6. Validated data will be fully documented to the required standards. Data will be made available to States via Web Site and/or STORET. Validated data will be ready 12-18 months after it is received from labs. This is the amount of time needed for all the data sets. Data sets can be made available as they are completed.

5. Outline of Data Flow for Coastal Group

1. Field crews from lead state agencies enter data into field computer system provided by the Southern California Coastal Water Research Project (SCCWRP). At this point, the state agency can use these data for whatever purpose they may have. If they choose to load these data to their local copy of STORET, they have to be sure not to upload them to the central STORET warehouse because that step will be done centrally by AED when all data files, including lab results and derived data, are complete.
2. Field crews send computer files to SCCWRP.
3. Field crews send samples (macroinvertebrates, chemistry) to analytical labs.
4. SCCWRP moves field data into Coastal database at SCCWRP.
5. SCCWRP makes field data available to all partners either by email or by putting them on the EMAP web server, with password protection.
6. SCCWRP starts preparing metadata files (with the help of the field crews).
7. Analytical labs send results back to whoever is the Project Officer for those samples; also sends a copy to SCCWRP.
8. Project Officer verifies that analytical lab work was done in accordance with contract requirements; notifies SCCWRP. This should be completed within 12 months of field collection (depends on contract language).
9. SCCWRP sends lab results to lead state agency (within 15 months of field collection).
10. SCCWRP incorporates lab results into Coastal database.
11. SCCWRP makes lab data available to all partners either by email or by putting them on the EMAP web server, with password protection.
12. SCCWRP continues metadata preparation.
13. Coastal analytical team (ORD, SCCWRP, Regions, states, tribes) analyzes data, using the Coastal database and develops indices and other derived data. If it would help data access, this Coastal database can be put on the EMAP internal web server, with password protection so that it can be used only by the Coastal analytical team. Otherwise, data files can be exchanged by email.

14. SCCWRP incorporates derived data into database.
15. SCCWRP finishes metadata files.
16. SCCWRP sends all data and metadata to AED to be put on the internal EMAP web site.
17. AED loads all data and data description information to the AED copy of STORET.
18. AED moves data and metadata from internal web site to public EMAP web site (within 24 months of field collection).
19. AED uploads data from AED copy of STORET to the central STORET warehouse (within X months of field collection; depends on completion of STORET enhancements; see Sect. 2). 40.

6. Outline of Data Flow for Landscape Group

1. Landscape group acquires necessary data (MRLC, streams, etc).
2. Compiles data, conducts QA
3. Analyzes data, landscape indicators
4. Produces landscape results (maps, atlas)
5. Prepares metadata files.
6. Environmental Sciences Division (ESD) sends all data and metadata to AED to be put on the internal EMAP web site (within 22 months of start time)
7. AED moves data and metadata from internal web site to public EMAP web site (within 24 months of start time).

Table 1. Data Flow in the EMAP Western Pilot Study

	Field Crews	Resource Data Centers	EMAP-IM	EPA Public Web Site
Functions	Field data collection, Data entry Initial QA	Data entry, verification, validation, compilation Documentation QA (database integrity) EMAP data analyses and reports Data distribution to partners	QA (Consistency checks, completeness) System development Data distribution to partners	Data distribution to all others
Contents	Data logger files Field paper forms Sample tracking	Resource group database Metadata Statistical programs GIS coverages	Data Directory Preliminary data & metadata GIS coverages Bibliography Reports	Data Directory QA'd Data & metadata GIS coverages Bibliography Reports
Data Format	Field computer system Data logger files Paper forms	Database system files (SAS, Oracle, Access) Arc/Info GIS files	ASCII data sets Oracle, SAS Arc/Info export files	ASCII data sets Oracle Arc/Info export files
QA completion	75%	90%	95%	100%
Users	Field crews QA officers	EMAP participants & partners (Region, state, ORD)	EMAP participants & partners	Other researchers & managers, academia, the public Other data systems: STORET, EIMS, GCMD
Time (months)	0	0 - 23	6 - 23	24

Fig. 1. Data Flow in the EMAP Western Pilot Study

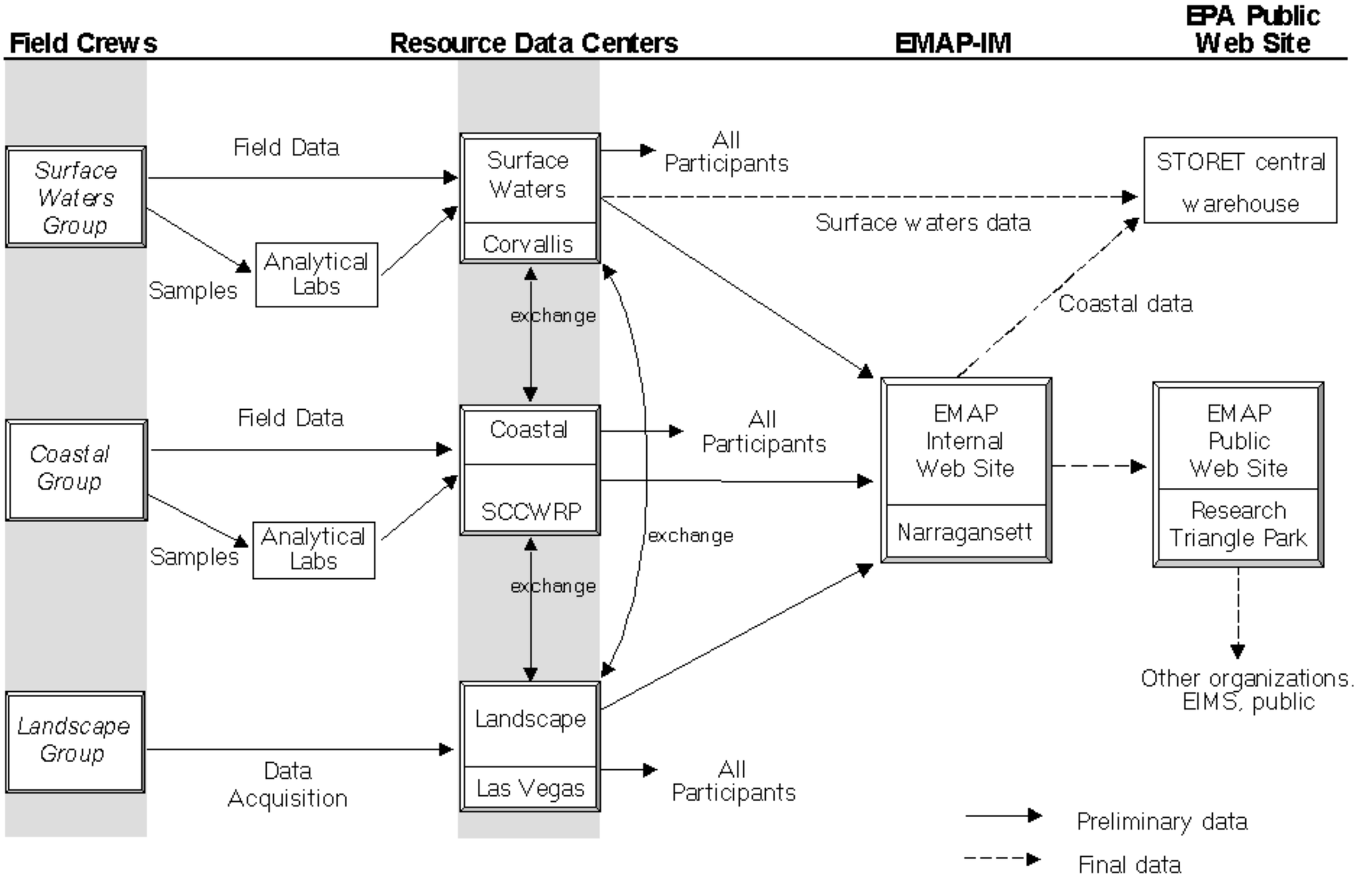

Fig. 2. Surface Waters Data Flow