U.S. Department of Transportation
Federal Highway Administration

CLARUS WEATHER SYSTEM DESIGN

DESIGN GAPS ANALYSIS

JANUARY 2006

Prepared By:

Mixon Hill

Mixon/Hill, Inc.
12980 Metcalf Ave, Suite 470
Overland Park, Kansas 66213
913-239-8400

Table of Contents

Table of Figures

Table of Tables

Revision History

Revision	Issue Date	Status	Authority	Comments
01.00	10/28/2005	Review	DTFH61-05-C-00022	Initial release for comment.
02.00	1/27/2006	Final	DTFH61-05-C-00022	Resolved comments.

Electronic File

Saved As: 04037-ak401gap0200.doc

1 INTRODUCTION

The purpose of the *Clarus* system is to collect, quality control, and disseminate environmental information. This design gaps analysis report identifies data collection, dissemination, and processing gaps, administrative and control gaps, and system infrastructural gaps for the *Clarus* system.

1.1 Purpose

The purpose of this document is to identify, assess, and provide solutions to gaps in systems supporting and interfacing with the *Clarus* system. This assessment will aid in bridging design gaps and building the detailed requirements for the *Clarus* system.

This document is intended to be read and used by the U.S. Department of Transportation (USDOT) and system development team members. As indicated in Figure 1, the "Gaps Analysis" is an intermediate deliverable in the larger context of the *Clarus* Weather System Design project, using criteria documented in the High-Level Requirements Specification to identify components and attributes ("potential features") as input to the Detailed System Requirements. This document subsumes and incorporates the "Solutions to Fill the Design Gaps" document shown in the figure.

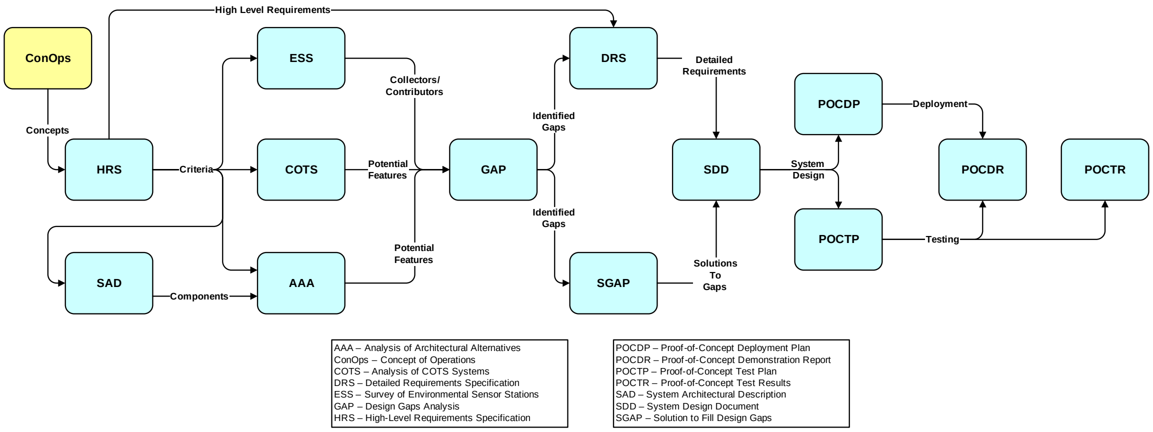

Figure 1 – *Clarus* System Documentation

1.2 Scope

This document provides an assessment of data collection, dissemination, and processing gaps, administrative and control gaps, and system infrastructural gaps that may limit the *Clarus* system's utility or acceptance by its stakeholders.

Clarus will collect weather and pavement[1] condition information from available environmental data sensors and mesonets. The system will qualify the environmental information using appropriate quality assessment methods to

---

[1] "Pavement" in this context includes surface and subsurface data specified in NTCIP 1204 (Ref. 9).

provide a relative indication of confidence in the information. *Clarus* will then format the qualified environmental information for dissemination to weather (or, more broadly speaking, environmental) service providers and for quality feedback to the environmental data contributors.

Clarus will provide benefits to a diverse group of stakeholders. Observing system owners and environmental equipment manufacturers will use *Clarus* information to improve the reliability and accuracy of their products. Transportation agencies will use qualified *Clarus* information to enhance their decision making in system operations and maintenance. Weather service providers, including the National Oceanic and Atmospheric Administration (NOAA), will use the qualified environmental information to enhance products distributed to the research community and the public.

1.3 Definitions, Acronyms, and Abbreviations

This document may contain terms, acronyms, and abbreviations that are unfamiliar to the reader. A glossary of these terms, acronyms, and abbreviations is provided in Appendix A.

1.4 References and Related Documents

1. *Clarus Final Draft Concept of Operations*; Iteris and Meridian Environmental Technology, Inc.; May 16, 2005.
2. *Clarus Weather System Design – High Level System Requirements Specification*; Mixon/Hill, Inc.; July 2005.
3. *Clarus Weather System Design – System Architectural Description;* Mixon/Hill, Inc.; August 2005.
4. *Clarus ESS Survey*; Cambridge Systematics, Inc.; September 2005.
5. *Clarus Weather System Design – Systems Engineering Analysis of Clarus-Related Systems;* Mixon/Hill, Inc. and Oklahoma Climatological Survey; September 2005.
6. *Clarus Weather System Design – Analysis of Architectural Alternatives*; Mixon/Hill, Inc.; September 2005.
7. *Environmental Sensor Stations owned by State Transportation Agencies*, summary compiled by Mitretek Systems, 2003.
8. *An End-to-End Quality Assurance System for the Modernized Coop Network*; Christopher A. Fiebrich, Renee A. McPherson, Clayton C. Fain, Jenifer R. Henslee, and Phillip D. Hurlbut; Oklahoma Climatological Survey, Norman, OK.
9. *NTCIP 1204 v02.23b NTCIP Environmental Sensor Station Interface Standard – Version 02*; National Electrical Manufacturers' Association, American Association of State Highway and Transportation Officials, and Institute of Transportation Engineers; 2005.

1.5 Overview

This document describes the potential technological deficiencies and inadequate coverage areas for the *Clarus* system. The design gaps for the *Clarus* system are documented in the Concept of Operations and further developed in the High-Level System Requirements, System Architectural Description, National Survey of ESS networks, Architecture Alternatives Analysis, and the Systems Engineering Analysis of *Clarus*-Related Systems (COTS Usability Report).

The remainder of this document consists of the following sections and content:

Section 2 Stakeholders and Concerns briefly summarizes stakeholders and their interests in *Clarus*, as more fully documented in the Concept of Operations.

Section 3 Operational Concept summarizes the Concept of Operations and identifies the major system elements.

Section 4 Design Gaps and Solutions describes the perceived design gaps and provides perspective on how the gap will be addressed by *Clarus*.

2 STAKEHOLDERS AND CONCERNS

The stakeholders in the *Clarus* system are identified and described in detail in the *Clarus* Concept of Operations (ConOps). The initial list of stakeholders whose user needs are considered in that reference includes:

- Observation system owners such as State Departments of Transportation (DOTs), municipalities, transit authorities, rail carriers, and universities;
- Instrument and observation platform vendors;
- Direct data users such as system owners and their contractors;
- National Oceanic and Atmospheric Administration (NOAA);
- Surface transportation weather service providers;
- General weather service providers;
- Research community; and
- Climate data warehouse and other non-surface weather interests.

The list further divides into those stakeholders who are providing data to *Clarus*, using data from *Clarus*, or both. Each of these groups has an interest in *Clarus* data specific to their particular organizational objectives and processes. The *Clarus* High Level System Requirements Specification documents these interests and provides the basis for determining what system features will be needed to meet the users' needs.

3 OPERATIONAL CONCEPT

The *Clarus* ConOps provides extensive discussions of the operational context, objectives, constraints, and system functions. These concepts are illustrated through discussion of an overall framework and operational scenarios for various user communities. Operational characteristics of the *Clarus* system itself are a subset of the overall framework and scenarios. The processes to be implemented in the *Clarus* system have been distilled from the framework in the ConOps and are shown in Figure 2 and Figure 3 below. This description focuses specifically on those functions to be fulfilled by the *Clarus* system and generalizes the interfaces based on the data types (rather than source types).

From the overall system perspective, the *Clarus* system will take in environmental data and metadata, and provide environmental metadata and qualified environmental data on request. The system will perform these operations based on data sharing agreements that define the terms of access and on quality control parameters used in assessing the incoming data. The system will need access to the environmental data networks and servers, and will need to provide network access for users requesting information.

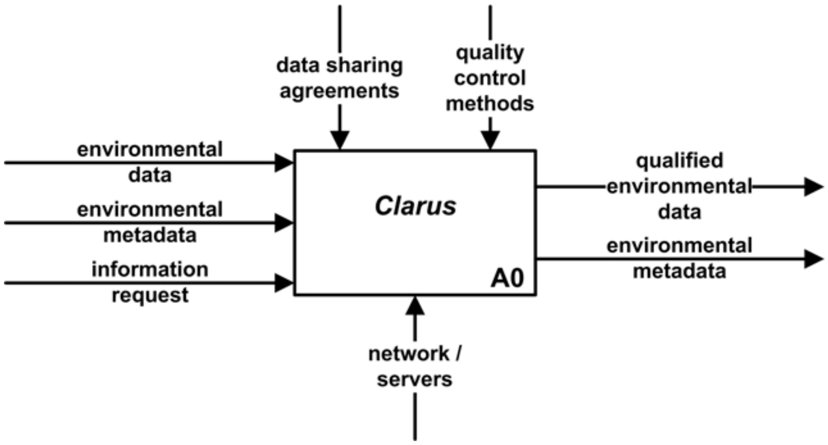

Figure 2 – *Clarus* System Process Context

Within the overall context, the *Clarus* system will collect, assess the quality of, and disseminate environmental data. The collection process locates, obtains, and stores the data in a common data structure, subject to access agreements. The quality control process applies one or more quality checks and associates quality flags with the data. The qualified data and the associated metadata are then available for dissemination, subject to any constraints specified in the data sharing agreements.

There will be multiple sources of data for the collection process, each potentially in its own format. Each source of data will also provide metadata describing the source and conditions surrounding the source. Terms under which data can be accessed from each source will be identified in data sharing agreements with the source organizations. Data are collected from the sources, interpreted from its source formats, and stored in a common data structure.

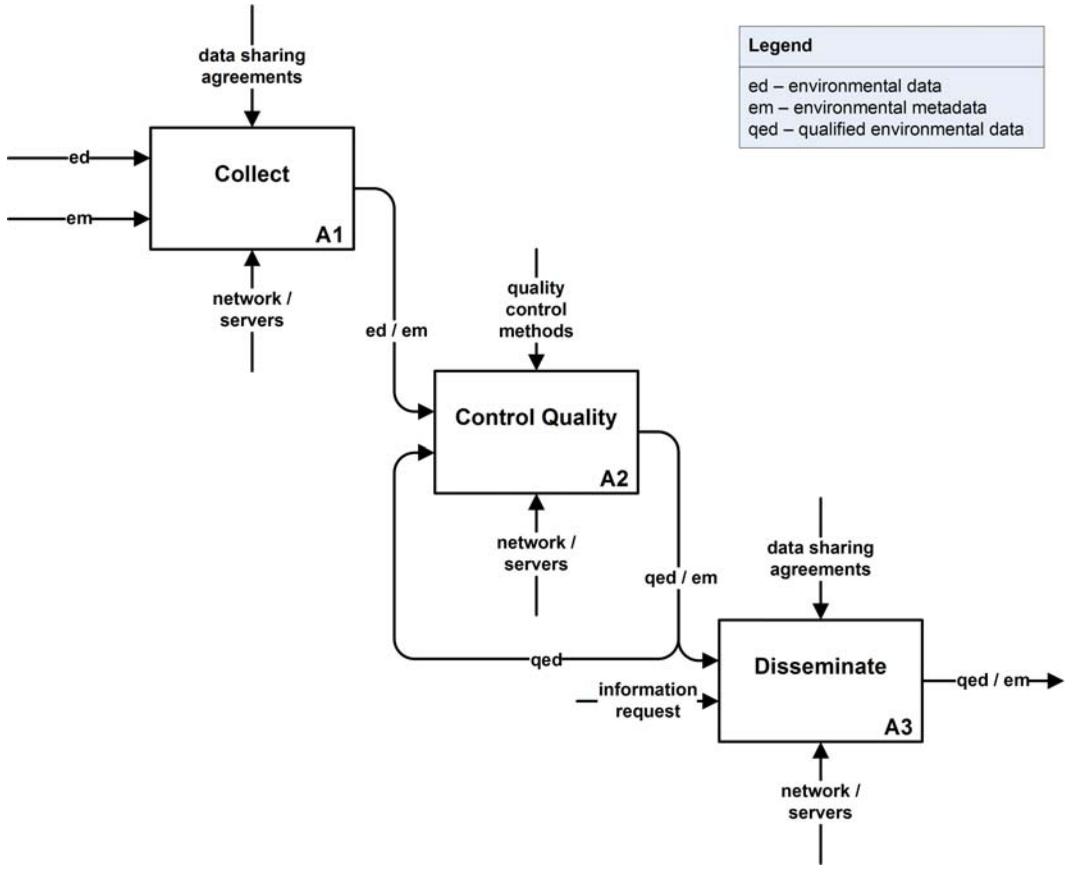

Figure 3 – High Level *Clarus* System Processes

The quality control process will implement one or more quality checks of the environmental data. Each quality check will be based on a set of rules for comparing the data to other models and data sets. Comparison data sets can include previously qualified data shown as a quality feedback loop in Figure 3. It may be necessary to derive or infer additional data from the original environmental data in order to complete some quality checks. Quality flags are assigned to the data according to the specific checks performed.

The data are disseminated in response to information requests directly from users with access to the data or automated processes based on data values and quality flags. In particular, data source organizations may be notified of data sets that do not pass particular quality checks. Data sharing agreements may constrain the data sets, formats, or distribution lists for data dissemination.

These essential *Clarus* functions provide the basis for developing the high level system architecture and design.

4 DESIGN GAPS AND SOLUTIONS

Generally speaking, design gaps are potential issues in the development, management, deployment, and operation of the *Clarus* system that may limit its ability to fulfill its intended purposes as described in the Concept of Operations and subsequent requirements specifications. Discussions in this section will identify, assess, and describe solutions to these gaps. Table 1 summarizes the design gaps and solutions.

Table 1 – Clarus Design Gaps

Gap	Issue(s)	Solution(s)
Data Sharing Agreements	• Protecting contributor interests • Enforcement imposes authentication requirements • Implementation complicates system design	• Use metadata to limit dissemination according to data sharing agreement
ESS Data Collection	• Low density of ESS deployment limits data availability • Uneven ESS distribution limits data utility	• Supplement with additional data sources such as VII
“Real Time” Data Access	• Delays in collector aggregation and distribution of data are outside *Clarus* control • Delays in *Clarus* processing	• Design system to reduce *Clarus*-imposed delay • Establish improved data reporting standards and encourage adoption
Data Collection Standards	• No industry standards • Multiple standards within an Agency’s individual network • Variation among a vendor’s hardware/software releases	• Design system to allow multiple file formats and protocols • Establish a *Clarus* interface specification for future ESS acquisitions
Metadata Reliability	• Existing metadata generally scarce and unreliable • No universal standards for metadata content and maintenance	• Include minimum metadata and maintenance requirements within data sharing agreement
Data Quality Checking Standards	• No universal standards for environmental data quality checking	• Design system to allow multiple methods • Develop *Clarus* data quality checking standard
Data Dissemination Standard	• No universal standards for environmental data formatting and reporting • Customized formats increase cost and complexity	• Provide data in simple universal formats (e.g., ASCII text)

4.1 Data Sharing Agreements

4.1.1 Gap

The effectiveness of the *Clarus* program and the usefulness of the *Clarus* system depend on making significant volumes of qualified environmental data available to a broad group of stakeholders. Since *Clarus* will not (at least in its initial phases) deploy its own environmental sensor stations and collectors, *Clarus* requires a broad network of data contributors in order to provide a consistent synoptic-scale view. The effectiveness of the data qualification methods likewise depends on access to a sufficient body of observations. *Clarus*' ultimate user satisfaction depends on being able to disseminate data provided by contributors to the broader group of *Clarus* stakeholders.

The terms according to which data are acquired will be controlled through data sharing agreements with the contributors. To be effective, the data sharing agreements must protect the contributor's interests without hindering the *Clarus* system's ability to use the data internally and to disseminate the data externally. In doing so, the agreements will be focused on the interchange of environmental data and metadata between the contributor and *Clarus*—but cannot arbitrarily constrain *Clarus*' ability to repackage the data for distribution.

Models for the data sharing agreement have been established for certain other environmental data collection networks. Among those systems evaluated in the *Systems Engineering Analysis of Clarus-Related Systems*, the Meteorological Assimilation Data Ingest System (MADIS) and the Canadian Road Weather Information Network (RWIN) system each utilize data sharing agreements to govern the collection and dissemination of ESS data.

The MADIS system utilizes three distribution categories or tiers to control user access to system data because some data providers have restrictions on the redistribution of their data. These tiers classify end users as either a NOAA organization, another government or educational organization, or all other parties. While this does address the issue of controlling data access, it imposes a user authentication requirement on the system that may not be consistent with other *Clarus* requirements. Further, these particular categorizations would prohibit value-added weather service providers from accessing all available data.

The Canadian RWIN system negotiates each data agreement individually, and then controls access to specific data streams based upon that agreement. The data remains the intellectual property of the Provincial DOT, so only internal distribution and viewing is allowed without permission from the system owner. The data can, however, be included in "derivative" production. In the context of *Clarus*, such a data sharing agreement would not induce a user authentication requirement into the system as does the MADIS model. Implementing this type of provision would, however, require defining terms such as "internal distribution and viewing" and "derivative" production. This approach will require more attention to the data sharing agreement drafts, but may provide the greatest number of users with access to the widest amount of data.

4.1.2 Potential Solution

To create consistency within the *Clarus* program, the terms of the data sharing agreements will be consistent between *Clarus* and all contributors. This will ensure consistent data coverage and accessibility across all contributor jurisdictions. The data sharing agreement will include provisions between *Clarus* and the contributor for:

- Communications;
- Quality checking standards for data and metadata;
- Coordination of activities;
- Intellectual property;
- Ownership and use of data;
- Indemnity and liability;
- Application of statutes, regulations, policies, and rules;
- Security;
- Other legal terms describing the relationship and roles; and
- Establishing ESS registration, specification, siting, descriptive metadata, telecommunications, maintenance, and configuration management standards.

With *consistent* data sharing agreements, the *Clarus* system is enabled to use metadata as the primary means of controlling distribution in accordance with those data sharing agreements. The metadata will be used to limit data dissemination by linking the contributor and user group identification to observation type (i.e. pavement condition). While this may impose additional processing demands on the *Clarus* system, it represents an acceptable compromise between system and user needs.

4.2 ESS Data Collection

4.2.1 Gap

Most state departments of transportation (DOT) have invested in Environmental Sensor Stations (ESS) along interstates and highways across the United States. The density of installations, manufacturer and type of ESS, and observation parameter lists vary greatly from state to state.

This section will address the gaps identified during the ESS Survey and their impact on the *Clarus* system if no modification or addition was implemented by state DOTs.

Table 2 shows statistical information about the number of ESS located across the nation. The table shows, by state, the number of DOT ESS, the total miles of interstate, principal, collectors, and local roadways, the area of land, the percent of highway miles based on the area of land, the number of ESS based on 1000 square miles of land, and the number of ESS based on 1000 highway miles.

Table 2 – State DOT ESS Statistics

State	Number of DOT ESS	Total Highway Miles	Area (sq. miles)	Highway Miles per sq. mile	ESS per 1000 sq. miles	ESS per 1000 Highway miles
Alabama	6	94,434	52,423	1.80	0.11	0.06
Alaska	56	14,230	656,425	0.02	0.09	3.94
Arizona	16	57,529	114,006	0.50	0.14	0.28
Arkansas	1	98,541	53,182	1.85	0.02	0.01
California	111	169,549	163,707	1.04	0.68	0.65
Colorado	83	86,821	104,100	0.83	0.80	0.96
Connecticut	9	21,089	5,544	3.80	1.62	0.43
Delaware	4	5,894	1,954	3.02	2.05	0.68
Florida	30	120,375	65,758	1.83	0.46	0.25
Georgia	61	116,534	59,441	1.96	1.03	0.52
Hawaii	0	4,309	10,932	0.39	0.00	0.00
Idaho	36	46,927	83,574	0.56	0.43	0.77
Illinois	87	138,526	57,918	2.39	1.50	0.63
Indiana	31	94,597	36,420	2.60	0.85	0.33
Iowa	53	113,516	56,276	2.02	0.94	0.47
Kansas	43	135,012	82,282	1.64	0.52	0.32
Kentucky	39	77,011	40,411	1.91	0.97	0.51
Louisiana	5	60,937	51,843	1.18	0.10	0.08
Maine	5	22,693	35,387	0.64	0.14	0.22

State	Number of DOT ESS	Total Highway Miles	Area (sq. miles)	Highway Miles per sq. mile	ESS per 1000 sq. miles	ESS per 1000 Highway miles
Maryland	63	30,688	12,407	2.47	5.08	2.05
Massachusetts	11	35,590	10,555	3.37	1.04	0.31
Michigan	33	122,222	96,810	1.26	0.34	0.27
Minnesota	93	131,893	86,943	1.52	1.07	0.71
Mississippi	0	74,105	48,434	1.53	0.00	0.00
Missouri	26	124,685	69,709	1.79	0.37	0.21
Montana	60	69,450	147,046	0.47	0.41	0.86
Nebraska	58	93,198	77,358	1.20	0.75	0.62
Nevada	70	33,977	110,567	0.31	0.63	2.06
New Hampshire	1	15,630	9,351	1.67	0.11	0.06
New Jersey	61	38,952	8,722	4.47	6.99	1.57
New Mexico	4	63,953	121,593	0.53	0.03	0.06
New York	63	113,124	54,475	2.08	1.16	0.56
North Carolina	34	102,160	53,821	1.90	0.63	0.33
North Dakota	20	86,782	70,704	1.23	0.28	0.23
Ohio	169	123,522	44,828	2.76	3.77	1.37
Oklahoma	2	112,578	69,903	1.61	0.03	0.02
Oregon	55	65,951	98,386	0.67	0.56	0.83
Pennsylvania	79	120,423	46,058	2.61	1.72	0.66
Rhode Island	15	6,415	1,545	4.15	9.71	2.34
South Carolina	24	66,230	32,007	2.07	0.75	0.36
South Dakota	40	83,688	77,121	1.09	0.52	0.48
Tennessee	81	88,518	42,146	2.10	1.92	0.92
Texas	71	301,987	268,601	1.12	0.26	0.24
Utah	60	42,716	84,904	0.50	0.71	1.40
Vermont	5	14,359	9,615	1.49	0.52	0.35
Virginia	41	71,242	42,769	1.67	0.96	0.58
Washington	84	82,264	71,303	1.15	1.18	1.02
West Virginia	6	36,993	24,231	1.53	0.25	0.16
Wisconsin	58	113,270	65,503	1.73	0.89	0.51
Wyoming	28	27,482	97,818	0.28	0.29	1.02
50 states	2091	3,972,571	3,786,816	1.05	0.55	0.53

The *Clarus* system is intended to create a nationwide surface transportation weather observing and forecasting system using DOT ESS to reduce the impact of adverse weather for road and transit users and operators. The DOT ESS survey conducted by Cambridge Systematics and the summary of ESS stations compiled by Mitretek Systems is represented in Figure 4 below.

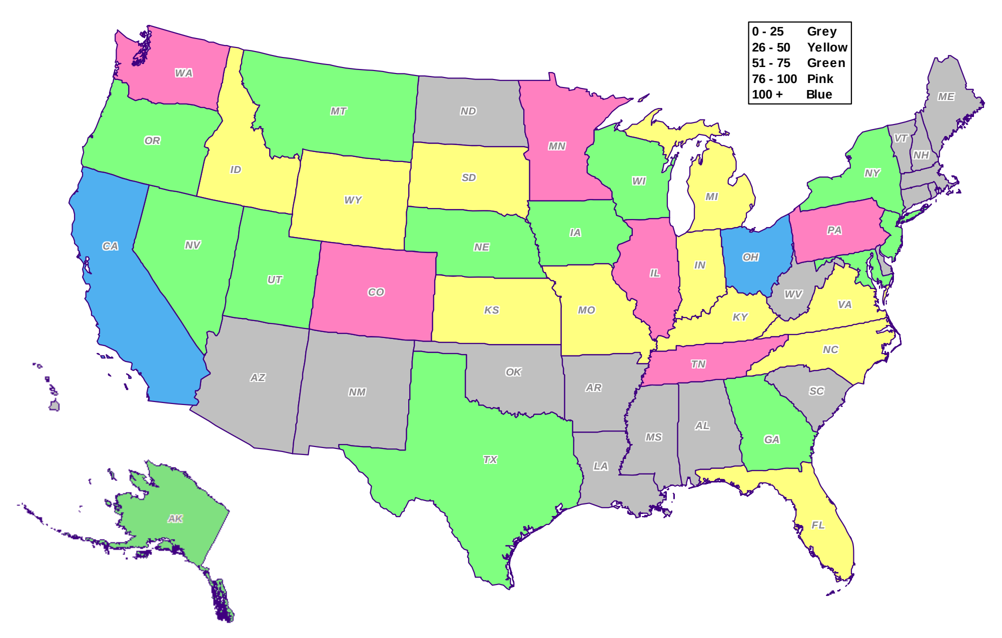

Figure 4 – National Coverage of DOT ESS

Based on this view of the national coverage, very few states have deployed a high number of instruments; this may preclude *Clarus* from meeting its full potential. Of the fifty states, only twenty-one states have more than fifty ESS installations.

Another view of ESS Coverage is depicted in the following Figure 5, showing each state’s coverage based on installations per 1000 square miles of land. Based on the low number of states with at least one ESS per 1000 square mile area (approximate mesoscale coverage), using the ESS data to build a mesoscale surface model becomes difficult without a sometimes substantial amount of data interpolation. As a result of the current distribution density of ESS across the states, the information from *Clarus* may be insufficient for state, county, and local agencies to make informed road management decisions.

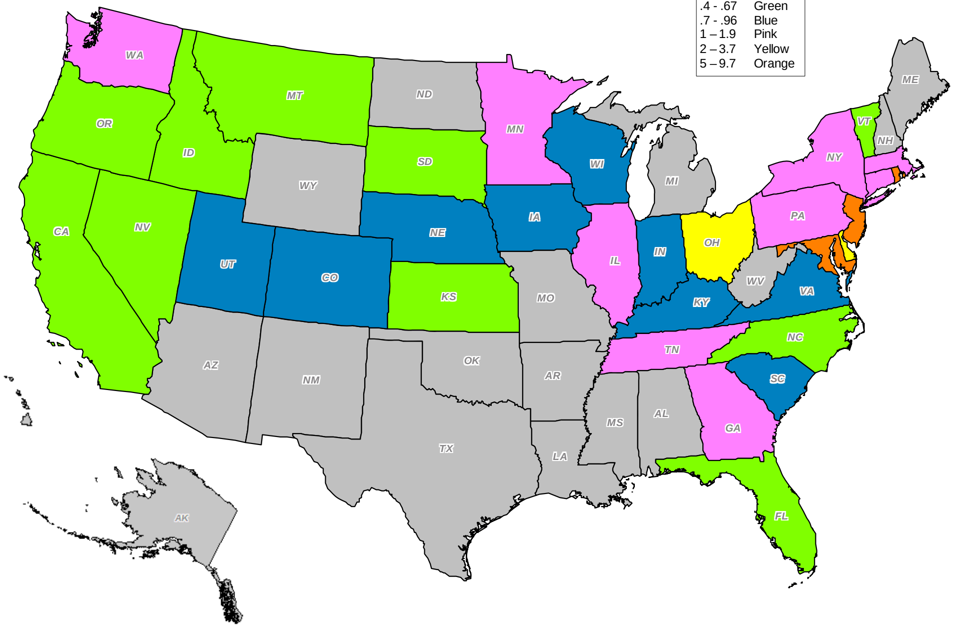

Figure 5 – 1000 Square Miles of ESS Coverage Based by State

Another view of the ESS coverage area is depicted in the following Figure 6. It represents the ESS coverage based on installations per 1000 roadway miles. Within each state, there are many different types of roadways – interstates, state primary and secondary highways, county roads, arterials, and streets.

Few states have sufficient number of DOT ESS to provide adequate coverage for their state's roadway network within the *Clarus* system. Recognizing there are 3,972,571 miles of roadways across the United States, if the existing ESS were distributed uniformly across the nation, there would be approximately .53 stations per 1000 highway miles. Naturally, ESS are not uniformly distributed across the highways either on a nationwide basis or even within a state. As shown in the following Figure 7 for example, some state DOT ESS are concentrated on bridges and metropolitan areas rather than distributed uniformly over the highway system.

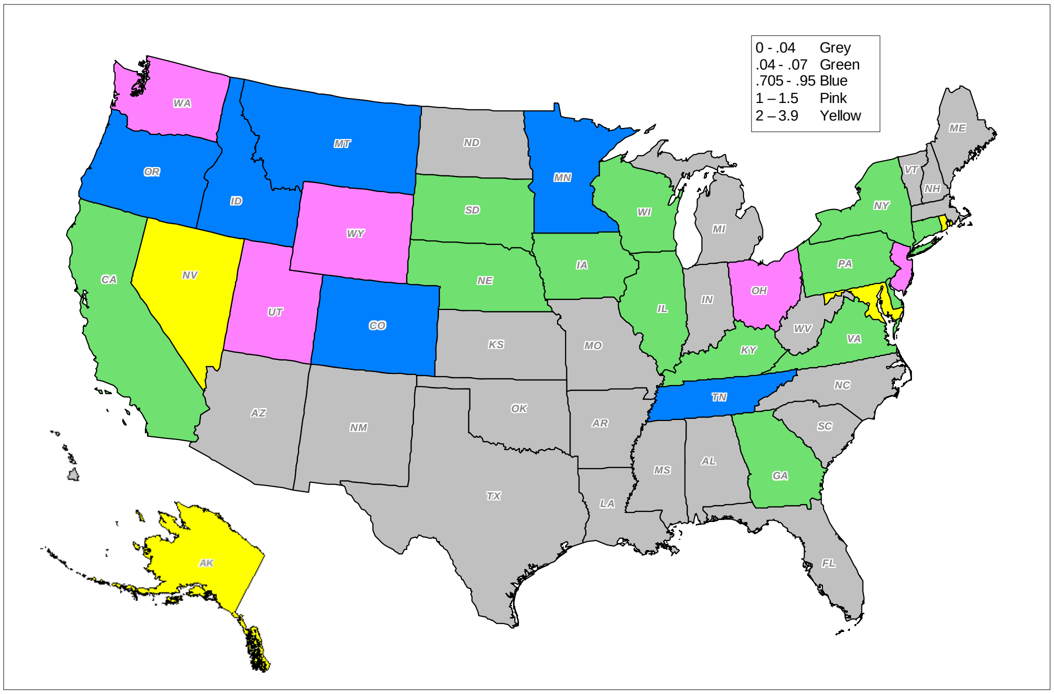

Figure 6 – 1000 Highway Miles of ESS Coverage Based by State

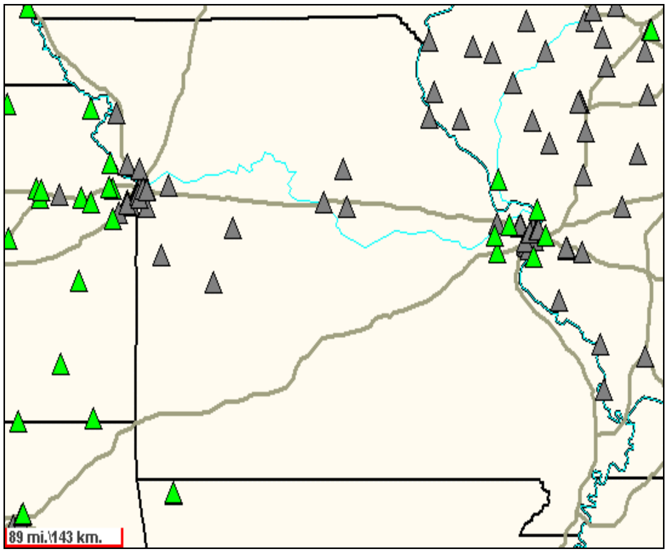

Figure 7 – Location of Missouri ESS

4.2.2 Potential Solution

While the number and distribution of ESS are not ideal for all desired analysis and forecasting applications, the *Clarus* system is designed and will be built to accommodate additional information sources. Scalability and flexibility are primary requirements of the system, allowing the system to adapt in both capacity and data collection standards. This means that data from other sources could be used to supplement the state DOT ESS data. For example, while supplemental data may initially be limited, ultimately *Clarus* is anticipated to ingest data from the Vehicle Infrastructure Integration (VII) initiative. These observations, when qualified, may address existing gaps caused by low quantity and uneven distribution of ESS installations.

4.3 "Real-Time" Data access

4.3.1 Gap

Timely access to environmental data is inherently valuable to surface transportation management and operations. Just as ongoing monitoring of traffic conditions enables operators to actively respond to congestion and incidents, having access to real-time environmental information enables operators to respond to developing adverse and emergency weather conditions. The benefit to operations lies in being able to make informed decisions that reduce injury and loss of life, damage to private property and the transportation infrastructure, and costs of avoidance and recovery.

As described in the high-level system requirements, *Clarus* users want to have real-time access to data from across their regions of operations. Ideally, data would be available to end users as soon as possible. In practice, this means making raw observations from *Clarus* available within fifteen-minutes of the actual observation time. Users seem to be willing to tolerate a small incremental delay for quality checking, but do not want unqualified observations to be withheld indefinitely.

As described in the ConOps, the total delay in making observations available from *Clarus* consists of time for:

- Collecting data from the ESS at the contributing agency's collector system;
- Collecting data from the contributing agency;
- Ingesting, quality checking, and publishing the data; and
- Creating any associated weather products.

Beyond the control of the *Clarus* system is the time delay between the ESS measurement and the owner agency's making the data available. This delay varies widely among collection networks, stations on those networks, and throughout the day and year. Underlying factors include the means and cost of communication with the ESS, capabilities of the collection network and nodes, and agency policies on data collection. These variations appear to be due as much to administrative or cost considerations as to technical issues.

The agency's systems also incur latency in publishing the data from the collectors. Even though the data may have been obtained from the sensor station in less than fifteen-minutes from the time of the observation, the collector system itself may not be publishing the data as soon as it is received. Some networks and collectors currently hold the data for release as a package of observations once an hour or once per day.

4.3.2 Potential Solution

As described above, the design gap is in obtaining access to data across all *Clarus* stations within the desired fifteen-minute interval. The *Clarus* design can minimize its own collection and processing intervals, but has no direct control over delays incurred in getting data from the field to the agency collector systems.

The *Clarus* system will be designed to minimize delays due to data processing. This includes designing and writing code that emphasizes processing efficiency as well as accuracy. In addition, the overall system architecture can be designed to distribute the computational and communication load at a physical location. Implementation could include clustered hardware at a particular site, a distributed architecture of regionally located *Clarus* servers, or a combination thereof.

Finally, the *Clarus* Initiative can encourage the use of improved standard reporting intervals. Through continued advocacy, the Federal Highway Administration (FHWA) can initiate a reduction in the current reporting interval, ultimately encouraging the state DOTs to reduce their reporting intervals to period that does not adversely impact the prompt distribution of data and helps the *Clarus* system meet the desired fifteen-minute interval. Data sharing agreements could provide a means of establishing these collection standards with the contributing agencies.

4.4 Data Collection Standards

4.4.1 Gap

Clarus, at least in the proof-of-concept deployment, will obtain data from collector systems, and not directly from ESS. While this network architecture reduces the number of nodes with which *Clarus* needs to communicate, it does not necessarily simplify the data collection process. These collector systems may each be designed to report observations in formats unique to that installation, even if using similar ESS equipment, network architectures, and data logging software systems.

In order to provide robust, reliable data services, *Clarus* needs consistent interface conditions at each of the data collectors. The attributes of the collectors needed by *Clarus* include:

- Means of access and network addressing;
- Levels of service (hours of service, bandwidth);
- Data object definitions (including units); and
- Data packaging, headers, and formats.

Operationally, *Clarus* will need notification of any changes to the collector attributes and, depending on the means of access, updates for application programming interfaces (APIs), drivers, and system tools.

The collector systems currently deployed have a variety of interfaces and formats. It is expected that the range of collector systems among the DOT road weather information system (RWIS) networks will include vendor-specific proprietary interfaces and DOT-specific solutions. Those DOTs with very few ESS may not have any collector systems at all, and therefore use direct ESS access to obtain observations. There is also likely to be variation among different hardware and software releases of similar systems from the same vendor.

Collector systems also have varying means of access and levels of service, potentially even among systems operated by a single agency. Many of the state DOTs, for example, operate some ESS themselves, and subscribe to third-party services for operating other ESS. In such an instance, *Clarus* will have to work with both the agency and the third party to have access to the full ESS installations.

4.4.2 Potential Solution

As discussed, *Clarus* will be required to operate with a variety of collector formats and protocols. This will be addressed by providing a *Clarus* collector service matching each format and protocol. Collector services will be matched to collectors in the system configuration data. It may also be possible to move toward consistency across the industry by defining and publishing a *Clarus*-standard format for future collector implementations.

4.5 Metadata Reliability

4.5.1 Gap

The availability of accurate and complete metadata is fundamental to the acceptance of *Clarus* data. Currently, significant gaps may exist in providing consistency in metadata accuracy and completeness, and in posting of changes in metadata.

In the most obvious case of acceptance issues, an observation is of limited value if the time and location of the observation are not available or accurate. More sophisticated analyses may require information about the means and circumstances of collecting, processing, and distributing the observation data. Inaccurate, corrupt, or out-of-date metadata compromises the usefulness of the entire system—not just the metadata that may be wrong.

Clarus needs accurate metadata to meet stakeholder expectations and assure its environmental data quality. Contributors will initially provide the metadata when a new collector system is added to the registry. Scope and accuracy will be subject to provisions of data sharing agreements. However, ongoing ESS and collector system operations and maintenance are likely to change metadata from time to time. Changes need to be provided to *Clarus* as these activities occur.

Clarus, in exchange, will provide feedback to contributors that may be helpful in operations and maintenance. For example, *Clarus* might determine that data were missing or were found, through quality checks, to be inconsistent with other observations and the presumed system state based on the metadata. The system owner could use this information to initiate or evaluate maintenance needs for the ESS or collector system.

Unfortunately, accuracy and completeness of metadata vary among contributors. Reporting and notification of changes likewise vary. Some of this variation occurs naturally as the consequence of seasonal variation in site conditions, drift in calibration, or aging of the sensor equipments. It can also be a consequence of constraints on the contributing agency's operations and maintenance budgets. Limitations in metadata change recording may be constrained by the system or organization managing the metadata. Whereas data reporting is automated, metadata management generally requires human intervention.

4.5.2 Potential Solution

For the *Clarus* system to fully realize its potential, accurate metadata is critical. To ensure the metadata remains accurate, regular metadata maintenance is required. Although the system will be designed to store extensive metadata, the metadata needed for the system to adequately collect and quality check observations will not be burdensome to the contributing agencies. The *Clarus* program can assist this process by incorporating metadata standards and maintenance terms within the data sharing agreements.

4.6 Data Quality Checking Standards

4.6.1 Gap

Quality control processes are the core value-adding services within the *Clarus* system. As described in the *Clarus* High-Level Requirements Specification, users of *Clarus* data will expect to get qualified data from *Clarus*, and to be able to identify the specific methods by which the data has been qualified. Contributors will expect to get notification that data may not have passed the quality checks, which enables them to take remedial action toward those collectors providing the data in question.

Some of the means for providing these quality check services are described in the *Clarus* ConOps and further specified in the high-level requirements. From that description and the requirements, it is expected that the quality check services will, for example:

- operate continuously;
- provide quality flags indicating the quality state of each observation;
- be able to implement rules for each environmental parameter;
- be able to implement rules for specific environmental situations (e.g., thunderstorms);
- be able to implement rules specific to observation locations;
- be able to base an algorithm on multiple observations;
- be able to use historical data; and
- be able to use multiple algorithms.

The high-level requirements do not, however, specify which quality control algorithms are to be used within *Clarus*. Although all stakeholders agree that *Clarus* needs to assess the quality of the incoming observation data, there has thus far been no focused effort or agreement on the methods. This is understandable since there are no particular standards for observation data quality assessment across the meteorological community. Finding methods for a common standard is difficult since different algorithms may be needed in particular regions or circumstances. Neither the meteorological research nor commercial communities are particularly motivated to establish common quality standards, since this would compromise the value of their competitive differentiators.

Other systems that gather environmental observations use a variety of quality control processing and reporting schemes. Descriptions of quality control methods in systems similar to *Clarus* were reviewed to gain a clearer understanding of the schemes being used. These systems, all of which were previously reviewed in the *Systems Engineering Analysis of Clarus-Related Systems*, included:

- MADIS;

- Canada’s RWIN, which uses the Canadian Meteorological Markup Language (CMML, version 2);
- NERON, NOAA’s Environmental Real-Time Observation Network; and
- MesoWest.

Generally, the quality processes in these systems are designed to generate a single quality flag value—either an alphabetic letter or single-digit number—that can be quickly and easily interpreted. The variability in quality flagging arises in reporting how the quality flag value was generated.

MADIS quality control uses letters to indicate a quality flag and categorizes algorithms into three levels of increasing scrutiny. Level 1 quality methods validate data using boundary condition values. Level 2 algorithms apply internal consistency and temporal consistency checks. Level 3 methods apply spatial consistency checks. The quality flags and their meanings are:

Z – Preliminary, no quality checking has been applied

C – Coarse pass, passed level 1

S – Screened, passed levels 1 and 2

V – Verified, passed level 1, 2, and 3

X – Rejected/erroneous, failed level 1

Q – Questioned, passed level 1 but failed level 2 or 3

Human oversight of the processes allows manual override of automatically generated flags. This process is called the “subjective analysis review” and uses “G” for a good evaluation and “B” for a bad evaluation. These characters, which are distinct from the automatic flags, indicate that the automated check was overridden by a person.

Two accompanying 10-bit fields are used to further categorize the type of algorithms applied and indicate failure of any one of the algorithms within a category. Five of the bits are reserved; the other bit position interpretation is as follows:

1 – Master check

2 – Validity check

4 – Internal consistency check

5 – Temporal consistency check

7 – Spatial consistency check

The master check bit in both bit fields is a roll-up value. It indicates if any check was performed and if any check performed failed. The remaining bits serve little purpose other than to further decompose level 2 algorithms into internal consistency and temporal consistency checks. For example, if a quality flag is “S,” “V,” or “Q,” it would be possible to inspect the bit fields to determine if internal, temporal, or both checks were applied. The reserved bits are probably intended for future definition and expansion.

MADIS documentation at http://www-sdd.fsl.noaa.gov/MADIS/madis_qc.html describes how the algorithms are applied and the scheme by which observations are included or excluded from the data set. No details were provided, however, on the names of or mathematical principles behind the algorithms typically used in MADIS.

RWIN and CMML are similar to MADIS in both summarized and detailed quality check specifications. CMML, being an XML derivative, defines a finite list of textual algorithm and quality flag values in its quality checking schema. The quality check types with their possible flag values are:

- presence
- integrity
- range
- temporal
- inter-variable
- spatial

- error
- missing
- doubtful
- inconsistency
- corrected
- estimated

Note that not all quality checking type and value combinations make sense. The overall quality summary information uses the same quality value flags with the addition of an “accepted/passed” value.

NERON, formerly known as NOAA’s cooperative observer program (COOP), has established mesoscale observer networks in the states of Maine, New Hampshire, Vermont, New York, Connecticut, Rhode Island, and Massachusetts. The automated quality checking system for NERON is provided by the Oklahoma Climatological Survey (OCS) and is based on their previous success with the Oklahoma Mesonet.

The NERON quality checking process first applies a filter to observations that allows a technician to flag the data as suspect, or that automatically flags an observation that has failed range tests or coincided with scheduled maintenance activities. Observations that have not been flagged are then processed by algorithms that perform spatial, step, persistence, and step-to-normal tests. These tests determine the final quality flag with the values:

0 – Good Observation;

1 – Suspect Observation;

2 – Warning; and

3 – Failure.

Suspect flag values are typically applied by the automatic algorithms, while warning flags are typically applied manually by technicians. Technicians review suspect or failed observations daily, often rerunning algorithms to determine how to correct the error. Sensor repair and replacement requests are generated to correct the problem when necessary.

MesoWest, operated by the University of Utah, gathers, processes, archives, integrates, and disseminates weather information from sensor stations across the western United States. MesoWest, like NERON, uses a single numeric value as a quality flag. The interpretation of their values is as follows:

-1 – Suspect: one or more of the variables were outside the acceptable ranges

0 – Unknown: flagging algorithms have not been applied to this data

1 – Caution: pressure, temperature, humidity, and/or dew point data should be used with caution

2 – OK: data has passed all defined quality control checks

9 – Missing: quality control checks beyond the max/min checks were not applied

All observations in the MesoWest system are flagged initially with a value of zero. The first quality checking algorithm, applied as the data arrives, compares the observation to an acceptable range. If the observation fails this test it is flagged as suspect (“-1”). Suspect observations are not used in subsequent tests and their flag value is not changed. Statistical regression analysis on 15-minute intervals over a six-hour period is then applied to the observation set. Observations are caution-flagged (“1”) based on the rules shown in Table 3.

Table 3 – MesoWest “Caution” Quality Rules

Observation Type	Rule
Temperature	Average temperature difference is greater than 10 degrees Fahrenheit.
Pressure	Average pressure difference is greater than 10 millibars.
Dew Point Temperature	Average dew point temperature difference is greater than 15 degrees Fahrenheit.
Wind speed	Difference between wind speed and wind gust speed is greater than 50 knots.

If a particular site did not report any observations for the previous six-hour period, then the quality flag is set to missing (“9”). Observations that are not flagged with values other than unknown (“0”) are then automatically set to OK (“2”). It is unclear from the available documentation if there is any manual intervention process used, but technician oversight is likely.

Although there are no universally-accepted standards for environmental data quality control, there have been some attempts to standardize quality attributes for geospatial data. The *Content Standard for Digital Geospatial Metadata* and its related ANSI NCITS 320-1998 description document do not deal directly with environmental observations, but it is possible to interpret environmental observations as geospatial attributes in this standard. These documents were reviewed for this analysis as a comparison of quality checking approaches in a closely related field.

The content standard does not codify the exact nature of all content, only the required and optional information types. Quality checking is largely performed by comparing coordinate sets to higher accuracy coordinate sets used as a standard reference. Content in this context is largely free-form text-based values with occasionally defined sets of statements and phrases called the domain.

The ANSI NCITS 320-1998 document requires geospatial data quality reports to include lineage, positional accuracy, attribute accuracy, logical consistency, and completeness. The lineage references the specific control information applied and lists and describes the mathematical transformations used in each step from the source coordinates to the final reported value. Positional accuracy descriptions consider the quality of the final geospatial product after all transformations are completed. There is also a manual input step with the positional accuracy descriptions that distinguish human deduced positional estimates from the results of other automated tests. Attribute accuracy is provided as a numerical estimated distance error value. Completeness information defines geometric thresholds such as minimum area or width. Completeness definitions also rely on standard taxonomy master lists such as Federal Information Processing Standards (FIPS) codes for states.

The system documentation reviewed herein reveals some information about the range of quality checking processes, but does not provide enough detail to identify any universal standards. The design gap in quality control processing lays in the lack of a standard naming convention and definition for mathematical quality checking algorithms. Quality checking algorithms can be complex; communicating which algorithms are applied, and in what order, is equally complicated. Algorithm-type categorization helps the situation for specific environmental data by allowing the back-end algorithms to be updated and replaced as needed without impacting the reported quality flags. This technique, however, obscures quality process details that may be important to a *Clarus* data user.

The gap is compounded by the inherent difficulty in interpreting a quality flag. A numeric-valued quality flag, for example, is easy to process and transmit. Higher-valued flags can easily be interpreted as being better than lower-valued flags. A simple value cannot, however, communicate the underlying detail. How much better is one quality flag value over another value? Which algorithms were applied? Which passed, and which failed?

4.6.2 Potential Solution

The *Clarus* program is effectively establishing its own standard set of quality-checking algorithms. The algorithms are being selected as current best practices from among those used in other existing environmental observation aggregation systems. Each algorithm is being reviewed by *Clarus* stakeholders, including meteorological interests outside the surface transportation community, for technical accuracy and applicability. Where applications of those methods vary among systems, the consensus technical opinion of the stakeholder community is being used to assure the broadest possible acceptance of the *Clarus* implementation.

Even if these standards aren't propagated outside the *Clarus* system context, the *Clarus* user community will know precisely what methods were being employed. Documenting the standards assists in the interpretation of the quality checking results without the need to track and provide detailed descriptions of the algorithms used for every possible configuration.

To further assure that *Clarus* quality checking services provide the best possible results, the system will be designed with adequate flexibility for multiple quality checking methods to be employed. The quality checking services will be extensible, allowing new methods to be implemented as they are proposed and agreed upon by the stakeholder community. Should it become advantageous, the system even offers the ability to supplement and tailor quality checking methods for particular observation types and locations.

4.7 Data Dissemination Standards

4.7.1 Gap

Existing observation publication systems use a wide variety of formats and protocols. Among the systems described in the *Systems Engineering Analysis of Clarus-Related Systems*, for example:

- MADIS provides multiple dissemination formats, including both binary formats by ftp and text formats through a Web browser.
- MesoWest currently provides graphical user interfaces for a broad range of stakeholders, with more limited binary and text formats.
- RWIN provides primarily text output in an XML format.

The variety of formats provided by the existing systems has presumably simplified user acceptance of those systems, but has not necessarily set a sustainable precedent for broader dissemination of environmental data. As the stakeholder community grows beyond meteorological professionals, it becomes critical that the dissemination formats become consistent and accessible, with minimal barriers to interpretation. The existing formats are tailored largely to forecasting professionals and the research community, and require a particular subject matter expertise to read. Users who are not meteorological professionals expect to be able to receive and interpret the data in human- and machine-readable formats based on common frames of reference without specialized tools.

Although the *Clarus* architecture allows for a set of output data and metadata dissemination services, it does not describe (nor do the high-level requirements specify) which formats are to be implemented in these services. More formats will drive up complexity and cost. Therefore, a gap exists between the variety of data dissemination formats used by current systems and the need to constrain the *Clarus* data dissemination formats to a subset of the universe of potential formats.

4.7.2 Potential Solution

The gap between the lack of universal dissemination standard and a practical resolution for *Clarus* can be met in part by establishing specific standards for *Clarus*-published data. *Clarus* will be best served by implementing the minimum number of standardized interfaces needed to serve its user community. These interfaces should be based on published formats used by a large fraction of the potential user community. Selected formats may be specific to particular user communities or industries (e.g., NetCDF for meteorological users) through a common protocol (e.g., HTTP).

Further, the *Clarus* data dissemination services will be configurable and extensible. This will facilitate implementing new *Clarus* data dissemination services as needed to accommodate future technologies and format changes.

APPENDIX A -DEFINITIONS, ACRONYMS, AND ABBREVIATIONS

The following table provides definitions of terms, acronyms, and abbreviations to assist interpretation of this document. *The IEEE Standard Dictionary of Electrical and Electronics Terms* [B2], IEEE Standard 610.12-1990, or IEEE/EIA Standard 12207.0-1996 may be referenced for terms not defined here.

Term	Definition
ANSI	American National Standards Institute
API	Application programming interface. A well-defined set of functions commonly used by software to interface with libraries of reusable algorithms.
Clarus	The *Clarus* system. An environmental data sharing system that collects, evaluates, and disseminates environmental data gathered from a geographically diverse set of environmental sensors.
CMML	Canadian Meteorological Markup Language.
Collector	An electronic device used to convert environmental sensor electrical signals into environmental condition measured values and store them for retrieval.
ConOps	Concept of Operations
Contributor	A managing agency or organization that owns and/or operates a set of environmental sensor collectors.
COOP	Cooperative Observer Program
COTS	Commercial off-the-shelf
DOT	Department of Transportation
ED	Environmental data. This data has not been processed by any quality checking algorithms.
EIA	Electronic Industries Association
EM	Environmental metadata. Information about an environmental sensor station.
ESS	Environmental Sensor Station
FHWA	Federal Highway Administration
FIPS	Federal Information Processing Standards
HTML	Hypertext Markup Language
HTTP	Hyper Text Transfer Protocol. A communication standard for transmitting and receiving documents and other types of data over the Internet.
IEEE	Institute of Electrical and Electronic Engineers, Inc.
MADIS	Meteorological Assimilation Data Ingest System
Metadata	In common information systems use, “metadata” is “data about data.” Within the meteorological community, this use has been extended to include data about objects related to weather observations. For example, location data for an ESS becomes metadata for the observation data.

Term	Definition
MHI	Mixon/Hill, Inc.
NCITS	National Committee on Information Technology Standards
NERON	NOAA's Environmental Real-Time Observation Network
NetCDF	Network Common Data Format
NOAA	National Oceanic and Atmospheric Administration. United States National Oceanic and Atmospheric Administration. A governmental administrative body responsible for managing programs and resources for weather and oceanographic science.
NTCIP	National Transportation Communications for ITS Protocol
OCS	Oklahoma Climatological Service
QED	Qualified Environmental Data. Environmental data that has been evaluated by quality checking algorithms and contains a quality assessment flag.
RWIN	Road Weather Information Network
RWIS	Road Weather Information System. Road Weather Information System. A unique system consisting of many meteorological stations strategically located alongside highways that allow the state Departments of Transportation to make more informed decisions during storms. Specialized equipment and computer programs monitor air and pavement temperature to make forecasts regarding how the weather impacts the operation and maintenance of the highways.
System	A collection of components organized to accomplish a specific function or set of functions.
USDOT	U. S. Department of Transportation
VII	Vehicle Infrastructure Integration
XML	eXtensible Markup Language. A flexible text markup language used to create standard information formats that share both the format and the information to enable the interchange of structured data.