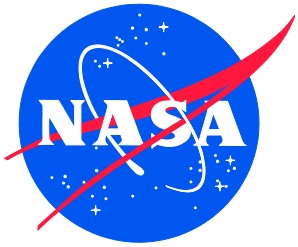
NASA/TM–2015-218991/Volume II
NESC-RP-14-00950



International Space Station (ISS) Anomalies Trending Study

Appendices

Robert J. Beil/NESC and Timothy K. Brady/NESC
Langley Research Center, Hampton, Virginia

Delmar C. Foster
Data Mining USA, Kennedy Space Center, Florida

Robert R. Graber
Science Applications International Corporation, Houston, Texas

Jane T. Malin
Johnson Space Center, Houston, Texas

Carroll G. Thornesbery
S&K Aerospace, Houston, Texas

David R. Throop
Jacobs Technology, Inc., Houston, Texas

December 2015

NASA STI Program . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counter-part of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI program, see the following:

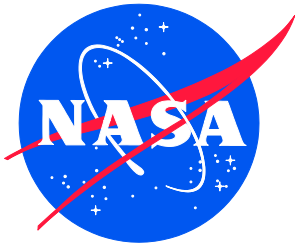
- Access the NASA STI program home page at *http://www.sti.nasa.gov*

- E-mail your question to help@sti.nasa.gov

- Phone the NASA STI Information Desk at 757-864-9658

- Write to:
  NASA STI Information Desk
  Mail Stop 148
  NASA Langley Research Center
  Hampton, VA 23681-2199

NASA/TM–2015-218991/Volume II
NESC-RP-14-00950



International Space Station (ISS) Anomalies Trending Study

Appendices

Robert J. Beil/NESC and Timothy K. Brady/NESC
Langley Research Center, Hampton, Virginia

Delmar C. Foster
Data Mining USA, Kennedy Space Center, Florida

Robert R. Graber
Science Applications International Corporation, Houston, Texas

Jane T. Malin
Johnson Space Center, Houston, Texas

Carroll G. Thornesbery
S&K Aerospace, Houston, Texas

David R. Throop
Jacobs Technology, Inc., Houston, Texas

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

December 2015

The use of trademarks or names of manufacturers in the report is for accurate reporting and does not constitute an official endorsement, either expressed or implied, of such products or manufacturers by the National Aeronautics and Space Administration.

Available from:

NASA STI Program / Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199
Fax: 757-864-6500

Volume II: Appendices

International Space Station (ISS) Anomalies Trending Study

September 24, 2015

Table of Contents

Appendix A. Outline of Concept of Operations (ConOps)—International Space Station (ISS) Anomalies Trending Study, December 9, 2014 3
Appendix B. Lexical Analysis of the Text in Anomaly Reports 23
Appendix C. Semi-Automated Ontology Updating from Corpus Analysis Results 27
Appendix D. Basic Process for Customizing and Updating the Aerospace Ontology 37
Appendix E. Data Visualization 46
Appendix F. Refining Proxy Codes 72
Appendix G. SAS® Analysis with Text-Mining Topics 82
Appendix H. ISS Data Mining Site Construction Guide 93

Appendix A. Outline of Concept of Operations (ConOps)—International Space Station (ISS) Anomalies Trending Study, December 9, 2014

A.1 Information for ISS Anomalies Trending Study Concept of Operations

The vision for the ISS anomalies trending study is to provide products to ISS discipline experts that are useful for analyzing ISS anomalies, starting early with immediately useful products and progressing to more capable products.

Initially, we expect to make heavier use of mediators (super-users) to direct use by discipline experts. We start by having super-users mediate the dialog between discipline experts and the combined, enhanced database. Data views are provided using Tableau®, a data visualization tool that offers multiple ways to graph and access data.[1] These views are constructed by super-users so that discipline experts can view and analyze the data. These super-users are team members who can observe the progression of analyses so the user interface can be tailored to fit those interactions. Gradually, we can transition to supporting a direct interaction between discipline experts and the data visualization software. This vision is supported by a three-phase delivery of anomaly data to discipline experts.

Table A-1 describes the database sources used in this project.

[1] Tableau® is described in detail at the Web site http://www.tableausoftware.com/.

Table A-1. ISS Data Sources

Source Type	Owners	Data Source	Acronym	Information	Contents Dates?	Nmbr of records	Comment
PART	Ames		Problem Analysis Resolution Tool	Ames data base to track ISS anomalies	1990s to present		Database is oriented to ISS.
		PART-IFI	Items for Investigation	Less detailed description of potential anomalies		?	become PRACAs. PRACAs are for problems that require more complex
		PART-PRACA	Problem Reporting and Corrective Actions	More detailed anomaly description		?	After a 2008 policy change, many problems that would have been reported as PRACAs were reported as IFIs to expedite their approval process.
GFE	QARC - JSC		Government Furnished Equipment - Quality Assurance Record Center (QARC)	data about NASA-developed (not contractor-developed) hardware and software	1993 to present		Database contains more than ISS, but we use only ISS records for GFE items.
		GFE-PRACA	Problem Reporting and Corrective Actions			11,000	Similar to PART-PRACA, but the database was designed later
		GFE-DR	Discrepancy Reports			220,000	
MOD	MOD - JSC	MOD-ARs	Directorate - Anomaly Reports		2001 to present	?	
Blended Data							Data available for reference, but not merged into a combined data base.
		MADS	Modeling and Analysis Data Set	Information on orbital replacement units (ORUs) and SRUs (one level below ORU)	?	?	
	multiple owners	State	Location in orbit, Increment, Status, Temperature	add to set as inquiries are made, timestamps are probably best way to link different info sets in STATE	?	?	
	MOD - JSC	SCR	Software Change Request	ISS software changes - response to bugs, anomalies, new feature requirements	?	?	contains new features as well as failures

A.2 Three Phases of Delivery

Figures A-2 through A-4 show three phases of delivery, moving from a quick start access of raw data to the full support for exploration by end users. Each phase of delivery is additive, so that more capabilities are added with each phase. Capabilities of early phases remain in later phases.

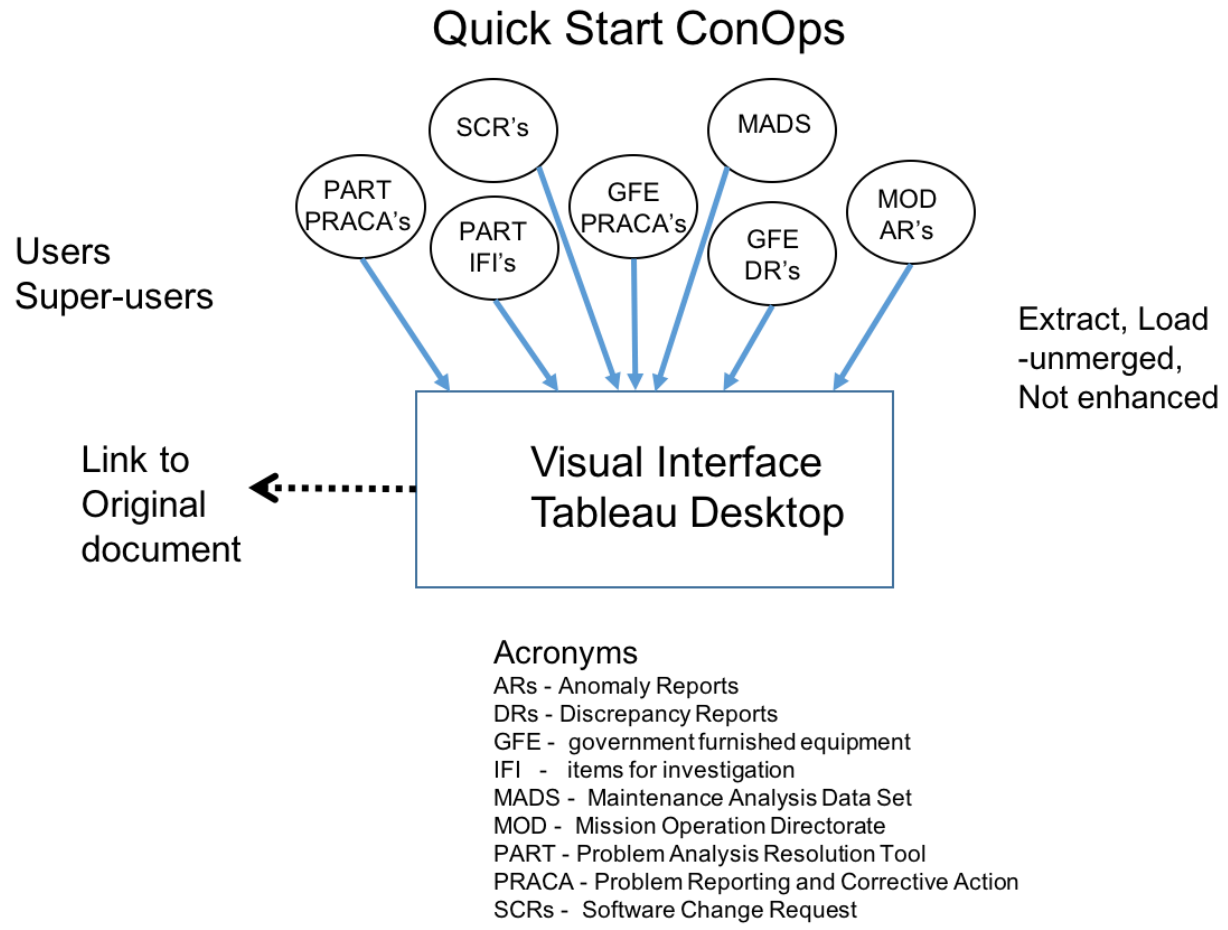


Figure A-2. Phase I, Quick Start ConOps, Within a Few Weeks of Starting the Project

Phase I, "Quick Start ConOps," provides access to combined data while data-merging decisions are still being worked out. This view can be created early and provides direct access to multiple data sources by discipline experts so they can more easily access the data. The views are created in Tableau® in a way that should be generally useful to discipline experts. They will support counts analyses of each database, enabling the identification of frequently occurring anomalies. As they use the data, they can request additional, more specialized views. Two levels of interactions for discipline experts will be supported: 1) basic, generic views of widespread

interest for identifying counts and trends of problem types and equipment, and 2) specialized views requested by specific discipline experts for follow-up investigations.

For this first phase, discipline experts are given access to a Microsoft® SharePoint® page with links to Tableau® software and to Tableau® data files containing information identified in Figure A-2. Super-users have constructed Tableau® views that allow flexible browsing of those data sources. These views allow users to see counts and trends of data, answering questions like "How many Problem Analysis Resolution Tool (PART) Problem Reporting and Corrective Action (PRACA) records have been coded with each failure mode?" and "Have the PRACA records with the failure mode of MA [mechanical assembly] been increasing over time?" These views also allow the user to navigate to the original records in the PART PRACA database. If the discipline experts have a need for views that are not already constructed, they can request special views, and the super-users will build them.

Phase 2, "Count ConOps" (see Figure A-3), provides access to merged data, with some supplemental data and a limited number of semantic tags. Phase II supports counts and trends analysis. The intent is to be able to look at counts and trends across multiple anomaly data sources. This requires the combining of data across data sources, which will require work to reconcile the way data are coded in each source. For example, one data source may have ten pre-specified cause codes, while another has 15 cause codes. This merged data coding scheme will need to accommodate data from both sources while retaining the information from each source. In another instance, cause codes may not be provided by a given source, in which case a proxy code will be added based on text mining of a description field that contains cause information.

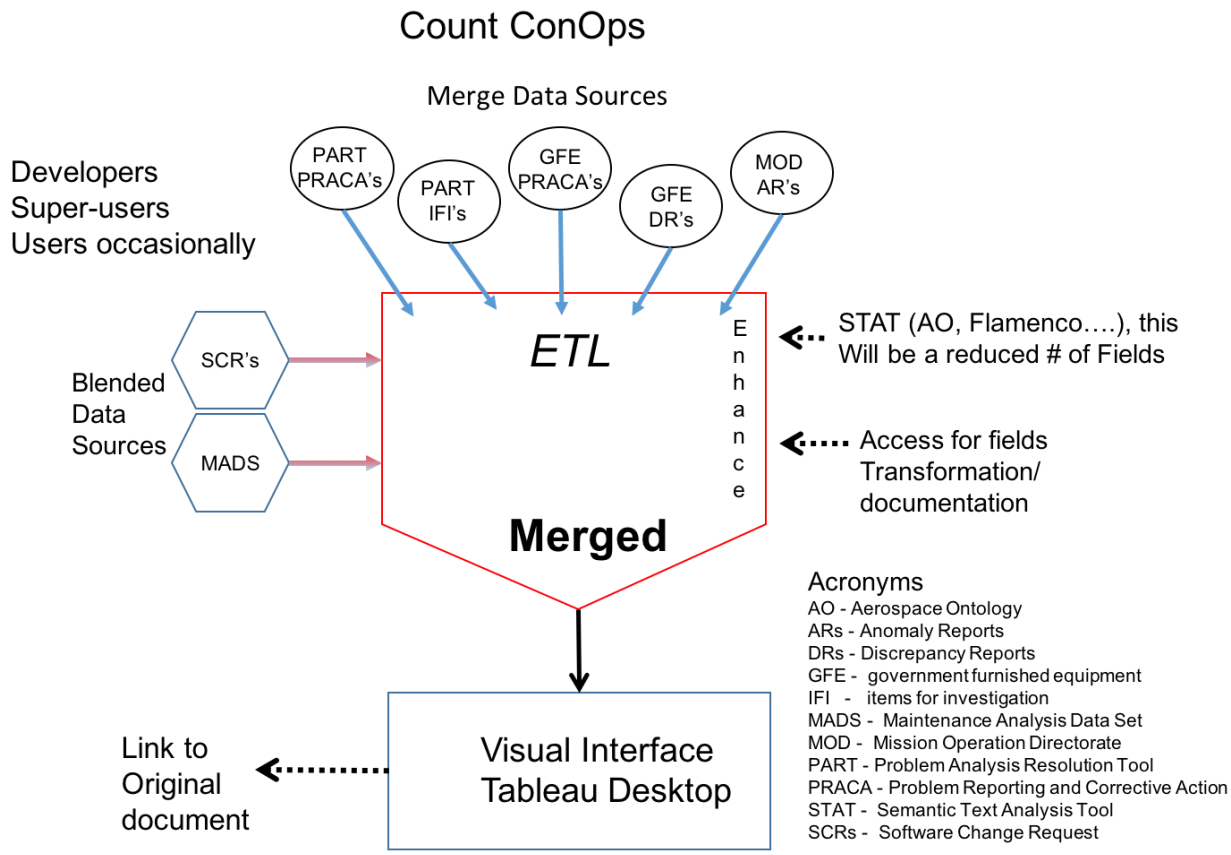


Figure A-3. Phase II, Count ConOps

Discipline experts will make use of the resulting database in a manner similar to that in the Quick Start phase. The quality of support for their analysis should be greatly enhanced by merging databases and by the proxy codes. In this case, users will only need to access a single data source that contains all the merged data (i.e., PART PRACA, PART items for investigation (IFIs), government-furnished equipment (GFE) PRACA, GFE Discrepancy Reports (DRs), and Mission Operations Directorate Anomaly Reports (MOD-ARs)). From that single access, users will be able to see all of the data regardless of its original source. While the names and contents of records from each data source are different, steps will be taken to make them similar to one another for viewing. A reduced set of the most informative fields will be selected, and fields with similar content will be given a common name for viewing purposes. Where possible, equivalent data values will be given a common label to make it possible to combine data counts across multiple data sources. Finally, proxy codes will be generated by data-mining software from descriptive text fields to supplement records that do not contain manually coded values for those fields. This should make it easier to look at counts and trends across all ISS anomaly-related data, regardless of the source.

The "blended" data shown in Figure A-3 are provided to enable further analyses of observed anomalies. For example, after discovering that there is a constant level of "power-on resets" of electrical switching equipment, a user may want to investigate possible causal factors or contributors to these events. After identifying the times when these resets occurred, the user could use those times to access state data that provide state information about the ISS. This could include whether the ISS is in direct sunlight, the temperature, and communications state. The blended data are not required to contain the same set of fields in the records as the merged data—the only requirement is that only some of the fields are common so that knowledge from the merged data can be used to investigate related blended data such as ISS state information.

A similar use of the blended data is illustrated by a user posing the follow-up question of whether the electrical switch problems are associated with the time they sent to the station. By starting with part numbers from the merged database, the user can look at the MADS database to when the switches were sent to the station.

Phase III, "Exploration ConOps" (see Figure A-4), supports a full exploration, integrating capabilities of semantic tagging and statistical text analysis. This capability is intended to take full advantage of semantic text mining and tagging based on the Aerospace Ontology and is presented for viewing dimensions compatible with the ways discipline experts need to view anomaly reports. Whereas the earlier phase was restricted to codes envisioned by designers of the component databases and proxies for those codes supplied by text mining, the exploration phase will consider browsing dimensions of the data that were not anticipated by database designers but would be useful to discipline experts in analyzing anomalies and risks. These additional browsing dimensions will be identified by exploring discipline-expert analysis targets implied in Section A.5 and by exploring options to address those analyses using the Phase 2 capabilities.

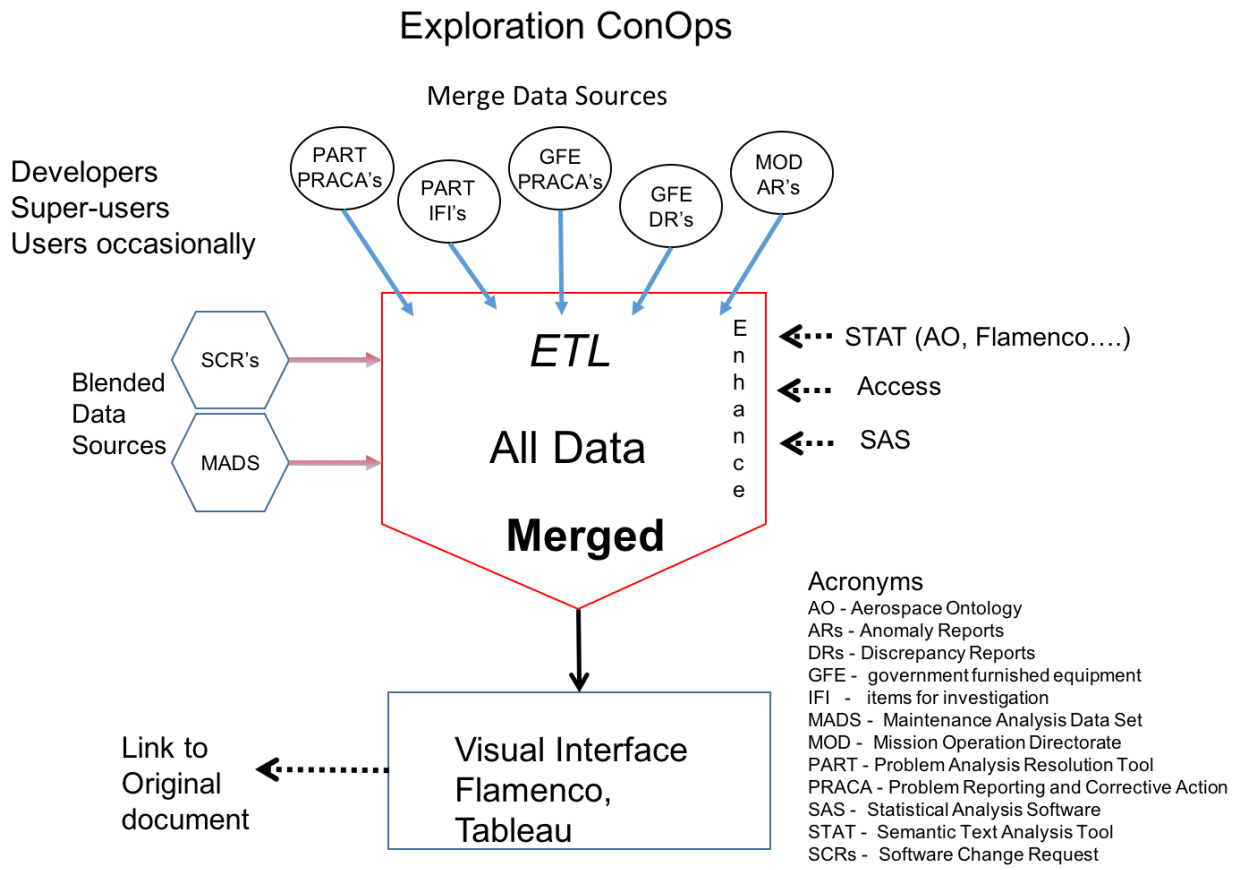


Figure A-4. Phase III, Exploration ConOps, near the Conclusion of the Project

Flamenco is a data visualization tool that has been used in the past to allow browsing of multiple, hierarchical semantic tags for anomaly data records. Flamenco is described in detail at the Web site http://flamenco.berkeley.edu/. We intend to explore the possibilities of deploying Flamenco for use by discipline experts or combining Flamenco output and Tableau® views. Useful data views are ones that help to answer the analysis questions in Section A.5 that discusses use case scenarios.

To accomplish this third phase of delivery, we anticipate the need for multiple capabilities of the team to exchange information in the manner illustrated in Figure A-5. This diagram shows how statistical and semantic mining efforts are integrated to develop an enhanced, combined database.

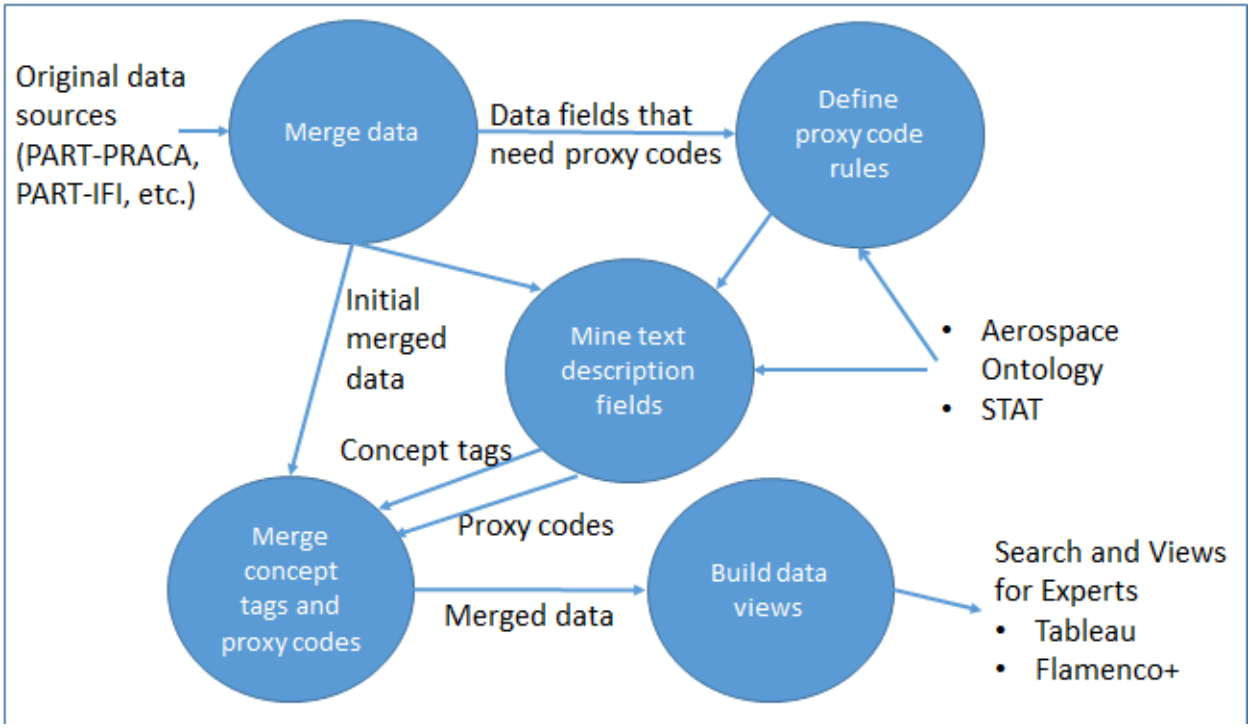


Figure A-5. Integration of Semantic Mining Efforts

Figure A-6 shows a product view of the capabilities for exploring ISS nonconformance reports. It shows the stages of transformation from the original data sources to the final merged data views, including enhanced search and visualization using Tableau® and Flamenco.

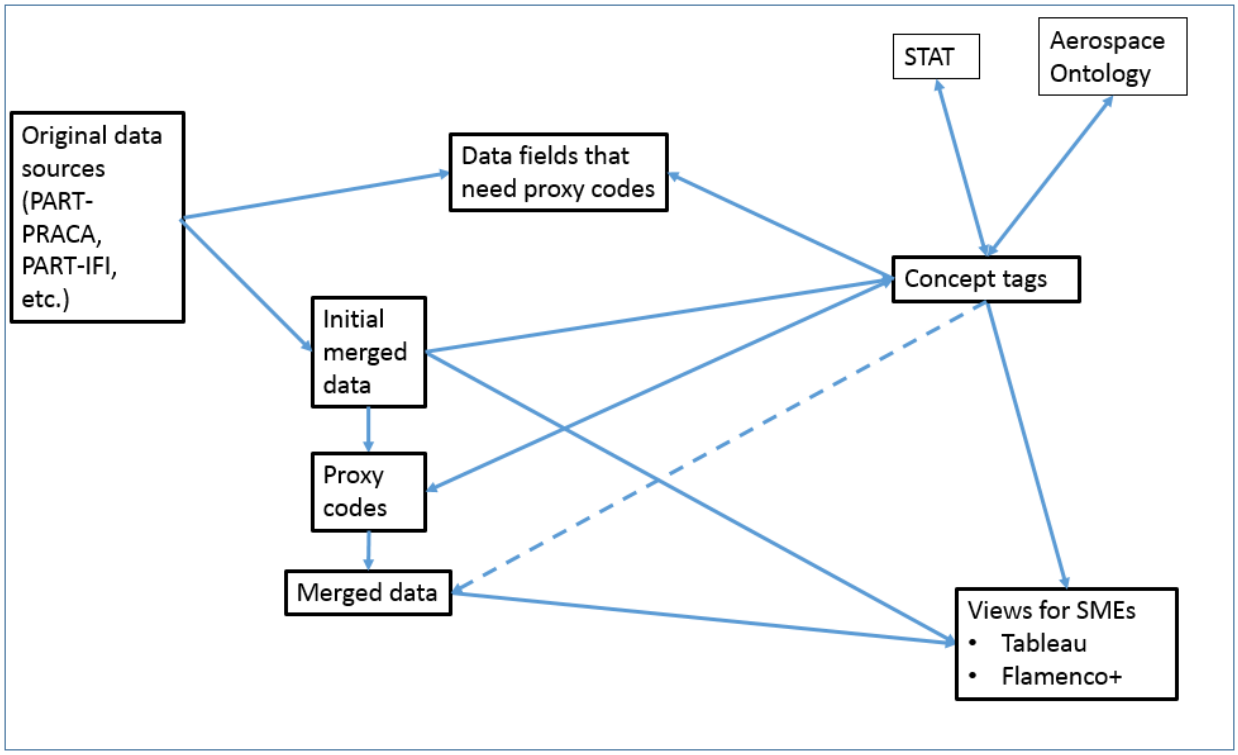


Figure A-6. Product View of Integrated Semantic Mining Efforts

Development methods for achieving the three phases of delivery are described in Section A.3. The anticipated dimensions for Phase III browsing are described in Section A.4, and the use case scenarios on which they are based are described in Section A.5.

In this third phase, "Exploration," the user should be able to not only see the merged views available in the second phase but also be able to browse the data in multiple hierarchical dimensions. For instance, a user might first look at the number of anomalies related to mechanical failure modes and whether they have increased over time. Then, the user might look at the relative numbers of mechanical failure modes for all the subcategories, and whether those related to hydraulics have been increasing. Later, the user may decide it is important to see what types of hydraulics issues are being observed (e.g., contamination, leakage, cavitation). Finally, the user may want to investigate how many contaminated hydraulics issues have involved a specific type of equipment. We anticipate using Flamenco to provide the capability of browsing along multiple hierarchical dimensions of the data in this manner.

The general interaction with discipline experts is illustrated in Figure A-6. When a new batch of anomaly data are extracted from the multiple source databases, super-users will build views to support most of the analyses that users will need. The exact nature of these views varies, depending on the phase of the delivery described above (i.e., Quick Start, Count, or Exploration). Users can then use those views to conduct their analyses. Occasionally, a follow-up question to

ask of the data will not be supported by the initial set of data views. If the required view is simple, the user may be able to construct the view; if not, the user can request a specialized view, which the super-users will construct. The new view will be used to address the follow-up questions. This process will continue until the users have enough information to complete a report on their analysis efforts.

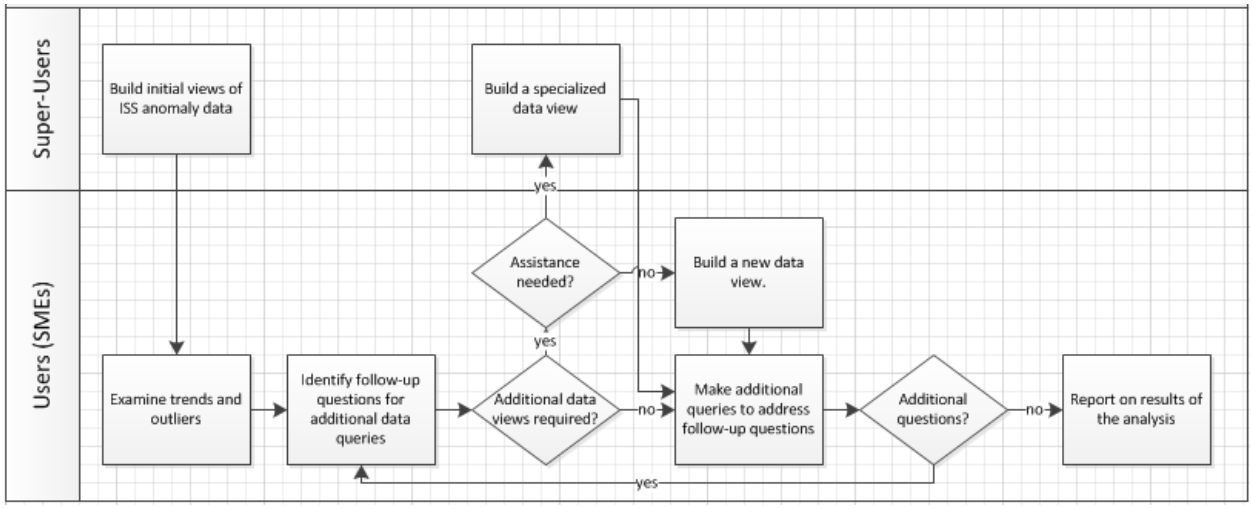


Figure A-6. General Interaction with Subject Matter Experts to Support Analysis of ISS Anomalies

A.3 Development Methods Supporting Browsing Dimensions

A.3.1 Merging Method (Phase I: Quick Start ConOps)

This is a description of providing access to a data viewer (e.g., Tableau®) and multiple sets of data from PART PRACA, PART IFIs, GFE PRACA, GFE DRs, and MOD ARs data sources.

- Go to each data source (e.g., DR, IFI, PRACA) and identify data fields informative for risk analysis and anomaly analysis. Use the data dictionary for each source.
- Build informative Tableau® views for each data source that allow discipline experts to explore the data and identify counts of anomalies from each source in the manner the data were coded by those who reported the anomalies (i.e., cause codes and failure mode codes as they were originally reported). Tableau® allows word search capabilities as well.
- All sources of data will be available starting from a single SharePoint® site.
- The Tableau® viewer used by discipline experts is free, with easy download instructions at the SharePoint® site.
- For most data sources, the discipline expert will be able to navigate to the original records from the Tableau® display.

A.3.2 Merging Method (Phase II: Count ConOps)

The purpose of the merging effort is to enable browsing of data merged from multiple sources so that all data for a given topic of interest (e.g., pump failures) can be retrieved regardless of the original data sources (e.g., DRs, PRACAs, IFIs). The challenge is that each source of data has a unique database structure. This method shows how they are merged:

- Start with the data fields from each source identified in Phase I as informative for risk analysis and anomaly analysis.
- Identify the right data fields for the merged data (important for accessing or understanding anomalies).
    - Combine similar data code fields from multiple sources. If the data codes identify a similar concept, then the codes probably should be combined (e.g., a "defect" code from one source might be essentially the same concept as a "problem" code from another source).
    - Keep data fields separate that describe different concepts. Sometimes data fields from different sources will have the same name but address a different concept (e.g., "status" from one data source may indicate a stage in a process flow, while from another source it may indicate whether a component was replaced).
- Identify the right set of data values for each of the merged data fields (important for accessing or understanding anomalies).
    - Combine data values from multiple sources that identify the same conceptual value. Some values from multiple sources will have different names but be essentially the same value. A good value name should be determined, and data from multiple sources should be assigned that value.
    - Keep data values separate that are conceptually different (e.g., "resolved" may not mean the same thing in different databases).
- Document the original sources of the merged data and value labels. Maintain a record of the merged data and how each data source contributes to the data. This allows the merged data to be traced back to the original record.

A.3.3 Tagging Method to Support Merging (Phase II: Count ConOps)

Some data fields do not exist in some data sources. For example, an anomaly report may not have a failure mode field. However, if a user is looking for all records related to a given failure mode, it would be helpful to see the anomaly reports that relate to the failure mode of interest. For this purpose, semantic analysis of text descriptions in the data record is used to generate "proxy codes" to stand in for the missing manual codes. This paragraph describes how "proxy codes" are added to make Phase II more useful to discipline experts.

Start with merged set (from Section A.3.2) to identify the target proxy codes for the semantic text mining to supply for data sources with missing data fields. Proxy codes are intended to identify what might have been entered by the person reporting an anomaly.

- If the person entered a manual code (e.g., cause code).
- If the person was permitted to report multiple codes for a given field (e.g., identify multiple causes).

The process for generating proxy codes follows:

- Identify merged data codes that need to be supplemented by combinations of tags identified by the Semantic Text Analysis Tool (STAT). STAT is an integrated toolset to analyze free text data fields to assign semantic tags that can be used to browse anomaly data like PRACA, IFIs, and DRs. These tags are associated with Aerospace Ontology concepts.
    - Some data sources will not contain reported codes for some of the merged data fields.
    - Use STAT to provide proxy codes for these records, where possible.
- Identify text description fields from each data source that can be mined for supplementing merged data codes.
- Map the ontology onto merged data codes (see Figure A-7).
    - Start with manual codes from data sources (e.g., cause codes).
    - This mapping involves the use of help text descriptions provided by database designers to help anomaly reporters describe the anomalies in a consistent, accurate manner.
    - Identify implied hierarchies for the coding levels.
        - For instance, a defect coding scheme may appear flat, with several one-, two-, three-, and four-letter codes, each of which has a help text description. However, looking at the codes and the help text (i.e., code definitions), an implied hierarchy can be detected. For example, several lines begin with an initial "E" in the code, and they are all electrical in nature. There are a few lines with an initial "EA," and these have to do with electrical assembly and installation. Four codes begin with "EAL," which have to do with electrical assembly and installation lead preparation. Hierarchical codes are illustrated in Figure A-8. Each level of these hierarchies needs to be mapped to parts of the Aerospace Ontology so that STAT can apply proxy codes compatible to those assigned by human anomaly reporters.
    - Run ATLAS routine from STAT against the help text for each code level. The ATLAS routine applies selected STAT capabilities without producing a fully

browsable database of anomalies. ATLAS provides views of STAT analysis components that are useful to developers but not to end users. It is applied to the help text to identify Aerospace Ontology concept tags. This allows an iterative testing and modifying of these capabilities and the Aerospace Ontology to achieve the desired tagging of anomalies.

- Using ATLAS output, manually identify matches between data codes and the Aerospace Ontology.
- Identify how to combine ontology concepts to form each proxy code. Some data codes may involve the combination of multiple parts of the ontology hierarchy to match the concept implied by the data code. For example, the defect code "DFH – Output Signal High" might involve the combination of the Aerospace Ontology concepts "Information_or_Signal_Object" and "Value_Above_Limit," as illustrated in Figure A-9.
- Since both STAT and the Aerospace Ontology are refined to reflect the desired tagging behavior for this help text, run STAT and use ATLAS to check how well ontology concept combinations form each proxy code.

- Vet the production of proxy codes by STAT. The Aerospace Ontology and STAT may require refinements, so this action may need to be done iteratively.
  - Run STAT to generate proxy codes from description fields from the merged data.
  - Compare STAT tags to manually entered codes where they exist.
  - Compare STAT tags to selected descriptive text to determine whether they look appropriate.

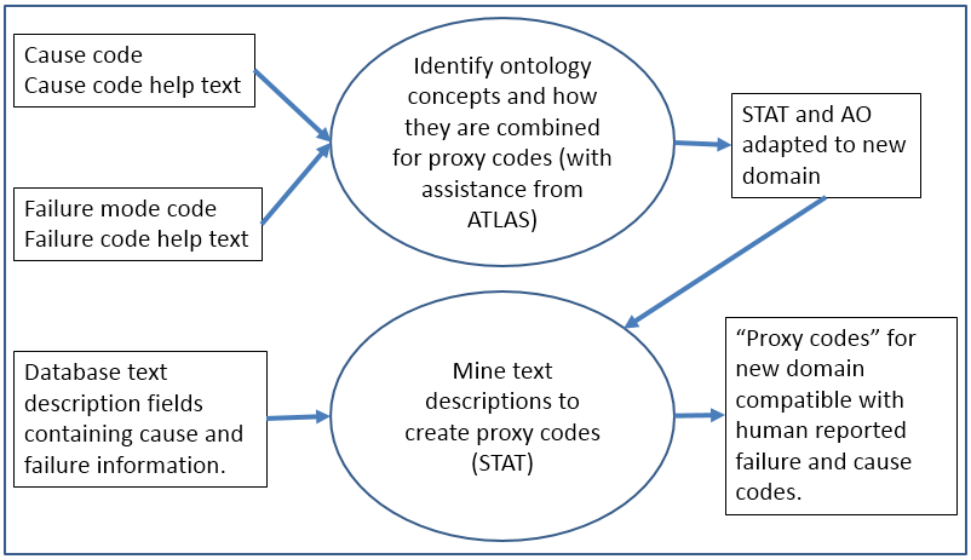


Figure A-7. Developing "Proxy Code" Capability in STAT

The relationship between developing proxy code mapping and using them to generate proxy codes is illustrated in Figure A-8. Figure A-8 calls out cause codes and failure mode codes in particular but could apply to all database codes of interest.

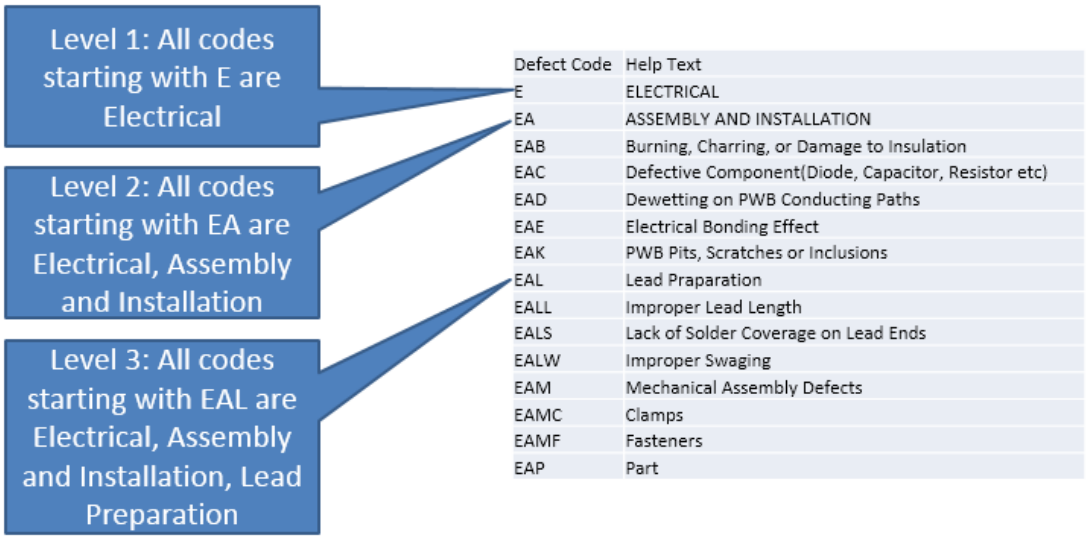


Figure A-8. Hierarchical Nature of Apparently Flat Database Codes

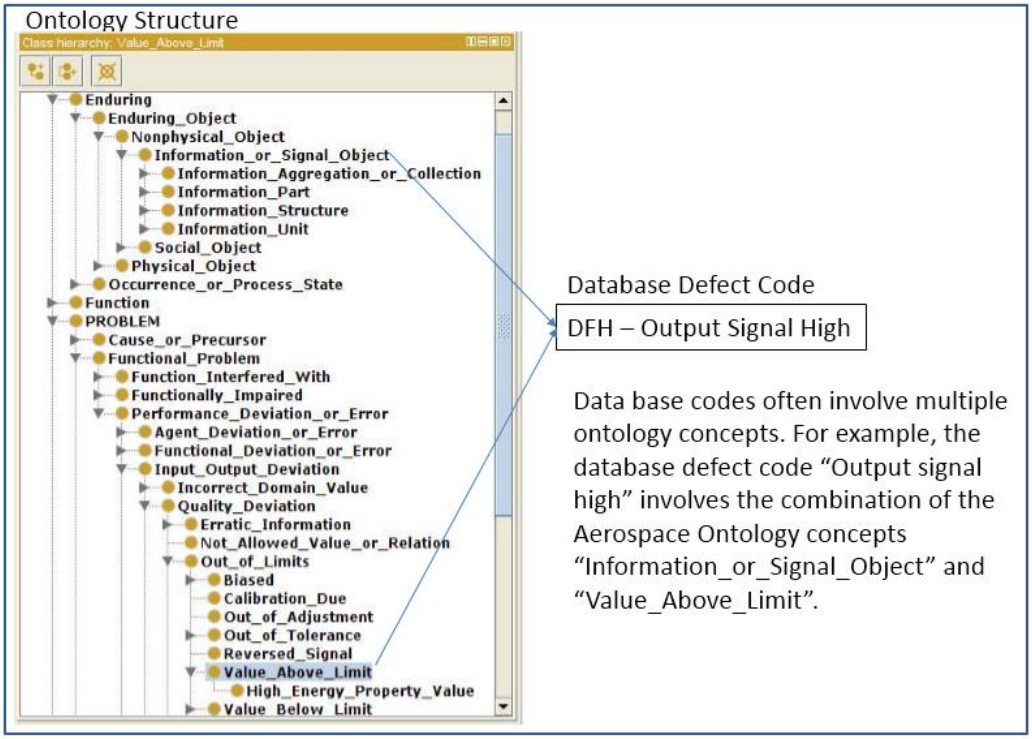


Figure A-9. Aerospace Ontology Concepts Often Need to be Combined to Form Proxy Codes

A.3.4 Tagging Method for Hierarchical Search and Browsing (Phase III: Exploration ConOps)

This last method is how to generate Phase III, "Exploration ConOps," to enable flexible browsing of anomaly data.

- Start with merged data from Phase II.
- Using the free-text description fields, create tags under browsing dimensions identified in Section A.4 to support analyses identified in expected usage scenarios.
- Include concatenated concept (topic) tag fields in the merged data set to enable the following scenario for using concept tags along with the remainder of the merged data fields to investigate issues in the ISS anomaly data. Combine use of concept tags, data base codes, and keywords to overcome search weaknesses.
  1. Perform a keyword search on words of interest for the issue at hand (e.g., "joint" AND "locking").
  2. Look at the resulting set of records from this search, with particular attention to the concepts in the concatenated concept tag field.
  3. Identify the concept tags that seem to define the issue at hand (e.g., "joint" and "mechanically impaired").
  4. Perform a new Tableau® search with those concept tags.
  5. Look at the resulting set for information related to the issue at hand.
  6. To further refine the search, if needed, look at the concept tags field to see how to refine the search and try again.
- Provide results in a browsing format that allows flexible browsing of tags in these dimensions. This may require the combinations of multiple data visualization capabilities like Tableau® and Flamenco. In Tableau®, the data set is the combination of all the data sources (i.e., GFE PRACA, PART PRACA, PART IFI, and MOD ARs). Using Flamenco for the first exploration of each data set allows the analyst to see what input data sources have the most information for further investigation in Tableau®.

A.4 Hierarchical Search and Browsing Dimensions

Browsing dimensions are intended to expose a combination of the data codes (e.g., defect codes) and the STAT tags (from semantic text analysis) so that the user can see problem reports that share a code regardless of the origin of that code (direct entry by the problem reporter, combination of tags from the data merging process, or tagging by STAT). The purpose of this outline is to identify useful ways for allowing users to browse anomaly data to support the analyses described in Section A.5.

Dimensions that are available from database codes include:

- ISS data source (from merged set, Phase II)

  - PART IFI
  - PART PRACA
  - GFE DR
  - GFE PRACA
  - MOD AR
- Response fields (from merged set)
  - Recurrence control
  - Disposition
  - Corrective action
- Environment (from merged set)
  - Increments
  - Ongoing activities (DR: prevailing condition; engineering activities; tests; test op code)
  - location: flight element
- Equipment (from merged set)
  - System – subsystem
  - Hardware level
  - Hardware type
  - Hardware category
- Time (time of anomaly)
  - Years (1995–2014)
    - Months (January through December)
  - Light/dark phases of orbit – solar angle
  - Equipment deployment times
  - Database entry rules (e.g., 2009 changes to allow MRB to close IFIs without making them into PRACAs)
  - Low versus high data periods

Hierarchical dimensions that are available because of STAT semantic tagging based on the Aerospace Ontology include:

- Equipment type
- Problem type
- Failure mode
- Defect
- Material
- Cause

A.5 Use Case Scenarios

These scenarios are in priority order. The names emphasize the overall analysis goals. Under each scenario are strategies for achieving that scenario and analysis questions that must be addressed.

Scenario 1: Counts and trends: identify recurring anomalies, emergent risks, recurring past precursors

- Identify counts (good, solid matches for accurate counts).
    - What types of anomalies occur most frequently?
    - What types of equipment experience anomalies frequently? Do the anomalies appear to be disproportionate for a part number or a vendor in particular, or do they appear to be related to the equipment type in general?
    - What are the top ten occurring problems in my discipline (e.g., thermal control)? What is the set of problem types that account for 80 percent of the problem reports in my discipline?
    - Are there new problem types or equipment types showing up in important numbers?
- Identify trends (good, solid matches for accurate counts).
    - Are some anomaly types increasing in frequency?
    - Are some equipment anomaly types increasing in frequency?
    - Are any problem types associated with the “big ten” on the rise?
    - Are these trends statistically reliable (e.g., Laplace Test)?
    - How many similar incidents should we expect in the future if no actions are taken (e.g., Crow-AMSAA test)?
- Identify outliers (source of follow-on questions for explaining the outliers).
    - What counts represent exceptions to trends, for example, one-quarter shows an exceptionally high number of problems?
    - What is the cause of this exception?
    - Find counts and trends within the exceptional category.
    - Identify environmental factors the might be related (e.g., flight increment, vehicles present at ISS, new deployments, ongoing anomalies).
    - See if the exception can be isolated to smaller subdivisions of any of the browsing dimensions (e.g., equipment, problem type, failure mode, defect, material, cause).
    - For example, what is responsible for the spike in electrical problems aboard ISS during the first quarter of 2010? What type are the counts of electrical problem types for that quarter? Is that a different proportion than for other quarters? What other important events occurred in that quarter?

- Conduct deeper analysis to generate candidate causes and corrective actions (e.g., broad search, accept higher risk of false positives so that we avoid missing relevant data).
    - Have we seen this problem type in the past? Is it related to a particular cause, equipment, subsystem, failure mode, or time interval?
    - Is this type of incident associated with a particular environmental factor? Increment? Light/dark phase? Are these features associated with a problem type?
    - Have we identified root causes, contributing factors, or other events that seem to occur just prior to this type of problem? Are any of these factors trending upward in frequency?
    - Have we identified root causes for this problem type? Has anyone made recommendations to address the root cause? Are those recommendations being followed?
    - What do failure modes and effects analyses (FMEAs) and hazard reports tell us about appropriate responses to this equipment-failure mode combination? What do FMEAs and hazard reports tell us about the possible consequences of this problem?
    - Having identified possible mitigations or preventive measures, does the body of anomaly reports have information regarding the effectiveness of these measures?

Scenario 2: Supporting an assessment

We have observed an incident that may be important.

- Have similar problem types occurred in the past? Are they increasing in frequency?
- Do they occur more in one location (e.g., flight element) or time (e.g., light/dark phase, high data interval, increment)?
- Are they associated with an equipment type, vendor, or model number? Is it very generic to equipment type or specific to a single model number?
- Are they associated with ongoing activities, prevailing conditions, or test activities?
- What corrective actions might be suggested for this type of problem?
- What additional consequences might we expect?

Scenario 3: Safety office evaluation

- Perform Safety and Mission Assurance (S&MA) assessments, evaluations, and studies to enhance the safety and success of programs and projects. Fiscal year (FY) 2014 ISS assessments on issues include:
    - Power-on reset anomalies.
    - Electrical power system high current oscillation anomalies.

  - Columbus interface heat exchanger close-call event.
- Similar analysis questions to those addressed in Scenario 2 above.
- Perform assessments and develop a comprehensive and integrated perspective of risk-based issues concerning vehicles supplied by NASA, international partners, and/or commercial entities. In FY14, "Quick Reference" guides to risk-based issues were developed for:
  - Russian vehicles
  - Automated transfer vehicle
  - SpaceX Dragon
  - Orbital Cygnus

The objective of these "Quick Reference" guides is to provide a comprehensive quick reference for decision makers at all safety reviews. These guides include information such as:

- Recent flight details of schedules, systems, configurations, and anomalies.
- Historical significant incidents and close calls.
- Spacecraft and launcher technical data.
- Launch, docking, undocking, and landing events and anomalies.

Scenario 4: Precursor analysis – If I can see a precursor, maybe I can predict and act on the problem before it develops.

- What anomaly types are occurring frequently enough to warrant evaluating them as precursors of events with possible severe consequences in the future?
- Which anomalies match concepts identified in hazard reports and FMEAs (e.g., failure mode, cause, controls, or effects)?
- How severe have the consequences of past occurrences of this anomaly type been?
- How frequent have the past occurrences of this anomaly been?
- Are they trending upward with time?
- Do system models (i.e., FMEAs, hazard reports) associate this type of anomaly with severe consequences? For instance, could a similar pump failure in another subsystem cause a loss of mission, vehicle, or crew?
- What is the frequency of occurrence of anomalies of other items in the causal chain identified in FEMAs and hazard reports (e.g., failure mode, cause, controls, or effects)? Are any of these trending upward?
- What additional equipment could exhibit a similar anomaly? What is the frequency and trending of anomalies for this additional equipment? Does this additional equipment have potential severe consequences according to FMEAs and hazard reports?

- For anomalies that were not considered to have a high enough risk value (i.e., likelihood and consequence) for a full quantitative risk analysis in the past, have they begun to occur at a higher frequency recently?

Appendix B. Lexical Analysis of the Text in Anomaly Reports

SAS® tools were used for lexical analysis of text fields in the anomaly report data. This analysis identifies words and phrases and the frequency of their use documents. Lexical analysis is a way to identify terms to be added to an ontology from documents in a new domain. SAS® Enterprise Guide was utilized to concatenate words in the text from fields in each data record, as illustrated in Figure B-1.

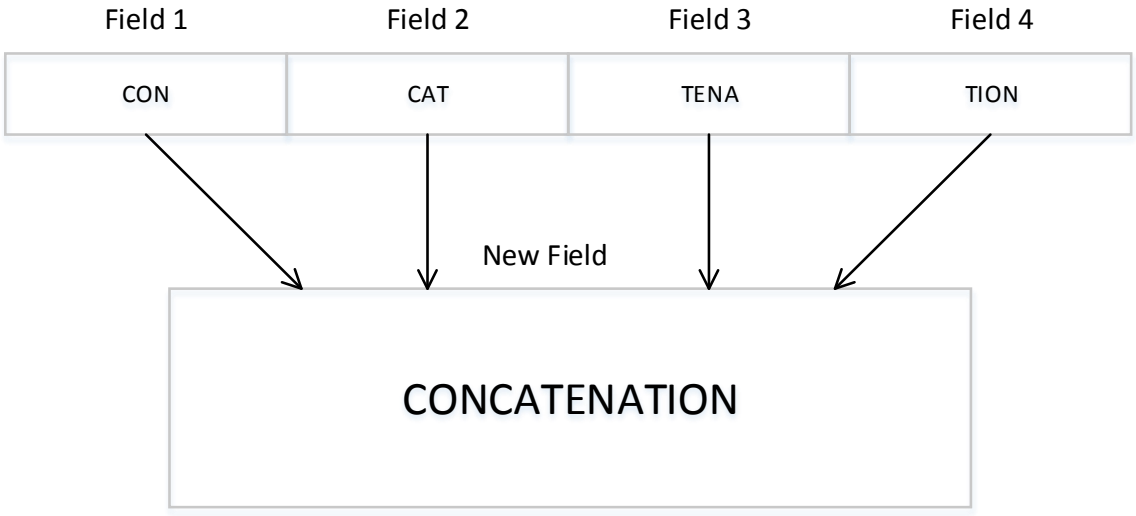


Figure B-1. Field Concatenation

Enterprise Miner and Text Miner were used to mine (lexically analyze) the combined data set, to find all terms and noun groups that might be added to the Aerospace Ontology. Some SAS® Text Mining nodes in the lexical analysis process are shown in Figure B-2. Approximately 170,000 different terms and noun groups were extracted from the 244,565 merged data records ("documents").

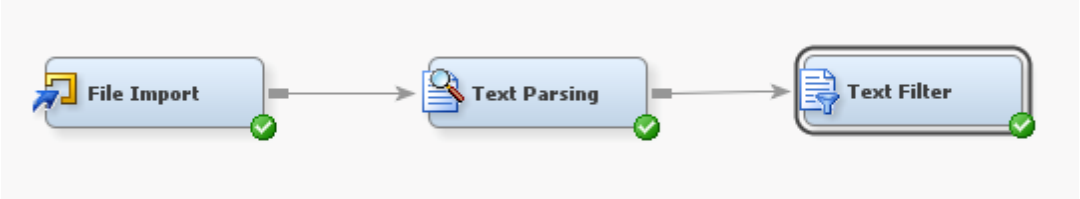


Figure B-2. Text Parsing and Text Filter Nodes

In lexical analysis, the Text Parsing node is most important. The Text Parsing node property sheet (see Figure B-3) shows properties for a typical analysis of Problem Description fields. SAS® files of engineering terms are used for some parts of this analysis.

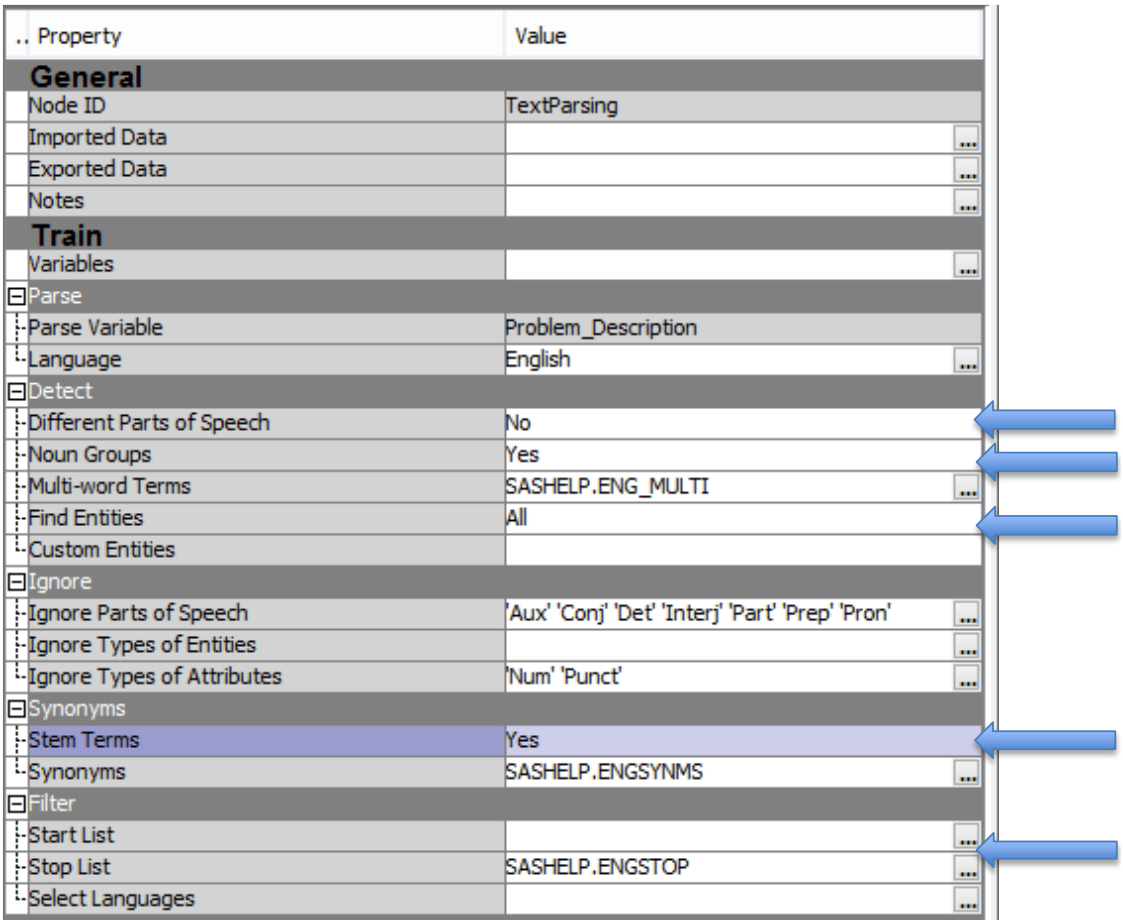


Figure B-3. Text Parsing Node Settings

Processing includes extraction of words, noun groups, entities, and multi-word terms.

- Noun Groups treat frequent term sequences as a single term (e.g., error message, ammonia leak, thermal cycle).
- Find Entities identifies sequences of characters such as phone numbers, names, and dates.

Stemming and Stop List filtering reduces the number of terms by eliminating redundant or uninformative terms.

- Stem Terms converts terms to their root form (e.g., “stems,” “stemmed,” and “stemming” all become “stem”).
- The Stop List excludes specified terms with low information such as “and,” “the,” and “is.” (The Start List includes specified terms during analysis.)
- Parts of Speech and Ignore Parts of Speech properties are used to identify nouns, verbs, adjectives, and adverbs. Knowing the part of speech distinguishes multiple meanings of terms with the same spelling. For example, see Figure B-4.

Term	Role	Freq	# Docs	Keep
excess	Adj	33	29	Y
excess	Noun	65	65	Y
excess	Verb	6	6	Y

Figure B-4. Different Parts of Speech

The Text Parsing and Text Filter nodes have data-cleaning features that help to remove relatively unimportant terms to simplify text mining. No further text filtering for noise reduction is needed in lexical analysis. The Text Filter weighting properties were set to default values, as shown in Figure B-5. To extract all the terms regardless of frequency in reports, the Minimum Number of Documents (anomaly report records) containing a term was set to 1.

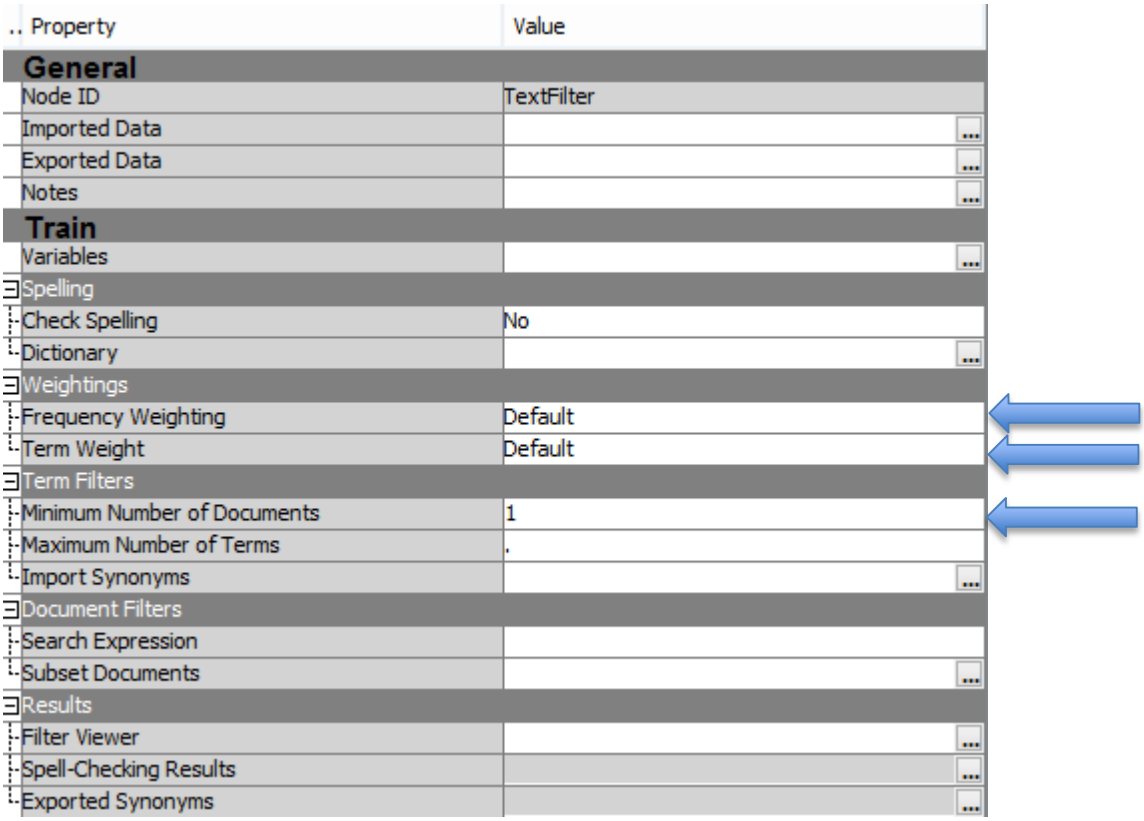


Figure B-5. Text Filter Node Settings

The process flow for lexical analysis is as follows.

- A set of documents (data records, in this case) were taken from problem reporting databases.
- The fields in the records were concatenated (as seen in Figure B-1).
- The text was parsed using the properties specified in Figure B-3.

- Extracted terms were entered into a (terms × documents) data matrix.
- A spreadsheet of terms and frequencies was the output.

Figure B-6 shows the data matrix in the context of a text mining process that includes a topic extraction phase. Topic extraction is discussed in Appendix G.

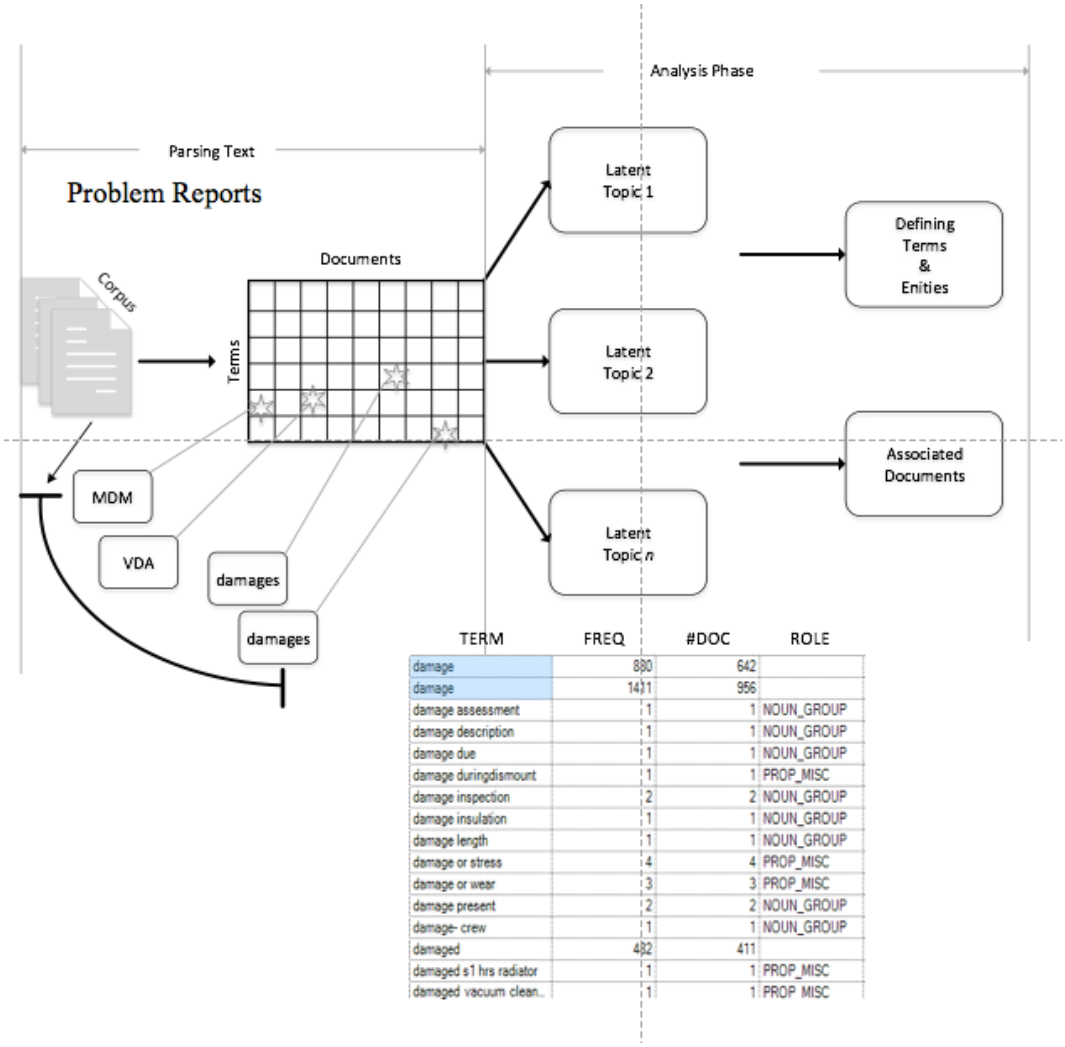


Figure B-6. Text Mining Processes and Lexical Analysis Output

Figure B-6 also shows part of an output spreadsheet that displays each term or noun group, with its frequency across all documents and the number of documents containing that term. The spreadsheet was used to identify terms and noun groups that might be missing in the Aerospace Ontology. Methods for using this output are discussed in Appendix D.

Appendix C. Semi-Automated Ontology Updating from Corpus Analysis Results

The Aerospace Ontology is the source of concepts (i.e., topics) used to match terms (i.e., words and phrases) identified in the free-form text fields of problem report data records by STAT (i.e., Semantic Text Analysis Tool). These concept-topics are used to enhance search, group records for displays of the faceted browsing application (Flamenco+), and generate and test rules for deriving proxies for manually designated defect codes and failure mode codes in government-furnished equipment (GFE) Problem Reporting and Corrective Action (PRACA) records. The Aerospace Ontology was developed during several previous projects, but the data sets in these projects did not include GFE PRACA or records from other databases in the merged data set.

The source of potentially important new terms was a large table of over 130,000 terms generated from lexical analysis of the text in the merged data set used in this assessment. The lexical analysis is described in Appendix B.

C.1 Reducing the Table of Terms

It is not practical to manually review 130,000 terms. A semi-automated method was developed to reduce the set and identify important new terms in the table that were missing from the Aerospace Ontology. This method led to selecting only 150 relevant terms, which was 0.12 percent of the original set.

The team developed software to clean the terms to remove numbers, proper nouns, and terms containing special characters. Long, multi-word phrases and phrases with embedded numbers were converted and eliminated. Rules for matching terms in the Aerospace Ontology were applied to the table of remaining terms. Tables of unmatched terms and matched terms were produced. After the first iteration of processing, the number of matched terms was about 54,200, and the number of unmatched terms was about 27,000. Table C-1 shows part of an unmatched terms table. Section C.5 provides a detailed description of the software processing.

Table C-1. Unmatched Terms

	A	B	C	D	E	F	G
1	TERM	Freq	Fail Word RL	Fail Word LR	Number of Words	Role	Reject
2	'WAS'	37665	WAS		1	Noun	
3	'WERE'	8601	WERE		1	Verb	
4	'HAS'	3855	HAS		1	Verb	
5	'OTHER'	3323	OTHER		1	Adj	
6	'ALSO'	2881	ALSO		1	Adv	
7	'DOES'	2819	DOES		1	Noun	
8	'BEEN'	2533	BEEN		1	Verb	
9	'RUSSIAN'	1492	RUSSIAN		1	Adj	
10	'SENT'	1384	SENT		1	Verb	
11	'RECURRENCE'	1320	RECURRENCE		1	Noun	
12	'BOTH'	1311	BOTH		1	Adj	
13	'SEEN'	1166	SEEN		1	Verb	
14	'ROOT'	1153	ROOT		1	Noun	

The terms in Table C-1 are ordered from highest to lowest frequency in the corpus. Terms can be words or phrases (i.e., noun groups). The "Role" column indicates part of speech or noun group. For words and multiword phrases (i.e., where the "Number of Words" value is greater than 1), the first word failing to match an Aerospace Ontology term in the search starting from the right is recorded in the "Fail Word RL" column. The first word failing to match in the search starting from the left is written to the "Fail Word LR" column if different from the "Fail Word RL." Fail words may be significant additions to the Aerospace Ontology. They would be difficult to pick out of the many multiword phrases without the "Fail Word" column information. The "Reject" column can be used to indicate terms considered but not included in the ontology update.

C.2 Review Strategies for Unmatched Terms

Terms were generally reviewed from most to least frequent. Sorting by frequency helps to focus the review on frequently used terms in the corpus. Frequent terms should be the most likely sources of material for updates to the Aerospace Ontology. Terms that can be associated with existing concepts or new concepts. The review strategy was to start with the 1,000 most frequent terms. A working spreadsheet of Aerospace Ontology additions was developed, with added columns to track terms selected from the unmatched term table and their frequencies. These extra columns were deleted from the version that was automatically imported into the Aerospace Ontology. A portion of this spreadsheet is shown in Table C-2.

Table C-2. Aerospace Ontology Additions Working Spreadsheet

	A	B	C	D	E	F	G	H	I
1	Class	Subclass	Creator	Comment	Member	Member	Member	Member	Member
2	Relative_Quantity	More_Than	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'ADDITIONAL'	2540	additional		
3	Description_or_Specificat	Document	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'CHIT'	1599	chit		
4	Record	Itemize	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'MANIFESTED'	1224	manifest		
5	Aerospace_System	Payload	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'MANIFESTED'	1224	manifest		
6	Position_Value	Inside	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'ONBOARD'	973	onboard		
7	Social_Artifact_Aggregatic	Nation	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'RUSSIAN'	867	nation	Russia	Japan
8	Achieve	Cause_Of	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'ROOT CAUSE'	769	root_cause		
9	Perform_or_Execute	Occur	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'RECURRENCE'	642	recur	reoccur	
10	Restore	Substitute	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'WORKAROUND'	488	workaround		
11	Location_Property	Position_or_Displace	Jmalin	abbreviation for degree. Source: lexical frequency analysis.	'DEG'	476	deg	degree	
12	Energy_Unit	Temperature_Unit	Jmalin	abbreviation for degree. Source: lexical frequency analysis.	'DEG'	476	degf	deg	degc
13	Time_of_Occurrence	Future	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'FUTURE'	476	future		
14	Information_Unit	Data	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'HEALTH'	467	health_status	health_bit	health_flag
15	Assertion_or_Fact	Quantifier	Jmalin	new class. Source: lexical frequency analysis.	'EACH'	466	all	some	many
16	Relative_Quantity	More_Than	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'ADDED'	460	added		
17	Performer_or_Agent_or_A	Operations_Agent	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'MERLIN'	455	merlin		
18	Process_Information	Ascertain	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'DISCOVERED'	454	discover		
19	Problem	General_Problem	Jmalin	Source: lexical frequency analysis of ISS problem report data sets:	'CONCERN'	447	concern		

For the first 1,000 terms, the frequency of selection generally decreased as frequency decreased. In the first 200 terms, 81 were selected. In the remaining 800 terms, 46 were selected. After the first 1,000, the review shifted focus to negative terms that might characterize problems (e.g., "difficulty" and "odor"). There were 23 terms selected from the next 4,100 terms. None of the selected terms appeared less than three times in the unmatched terms table. After review, about 150 terms were selected as the basis for adding new terms to the Aerospace Ontology. This is about 0.12 percent of the original set. For each selected term, one or more members of concepts were added. Less than ten new concept classes were added.

Many of the most frequent terms were easily rejected because they could have been stop words: general verbs or adjectives. Terms could be rejected if their stemmed roots matched words in the ontology. For example, "slow" is the root of "slowly" and "compliance" and "compliant" have the same root. Likewise, the root of a frequent term like "manifested" can be the version chosen to add to the Aerospace Ontology, as is shown in rows 3 and 4 of Table C-2.

Words can have different meanings in the context of noun group phrases. These phrases can be found in the table. Although these phrases are less frequent in the corpus, they indicate multiple meanings that should be included in additions to the Aerospace Ontology. Terms like "solar," "serial," or "health" are included in numerous phrases in various contexts in the table. For example, "solar flare" uses "solar" in a term associated with one concept (i.e., radiation), while "solar angle" uses "solar" in a term associated with another concept (i.e., property of bearing or orientation or pointing). Both can be included in additions to the Aerospace Ontology, thus pulling in terms that are less frequent but relevant.

Common misspellings can be found in the table (e.g., "recurrence" and "recurrance"). These can be added to an Aerospace Ontology concept that includes the correctly spelled term. Likewise,

other versions of terms can be found and added to the Aerospace Ontology. For example, there are five versions of deberthing in the table: deberthing, deberth, unberth, de berth, and de berthed. In addition, adding a term like "Russia" to a concept (e.g., nation) can lead to adding other terms (e.g., members) that are names of aerospace partner nations.

C.3 Reviewing Matched Terms

Table C-3 shows part of a matched terms table. Section C.5 provides more detail concerning the processing used to develop this table.

Table C-3. Matched Terms

	A	B	C	D	E	F	G
1	TERM	Freq	Matched Sequences	Match Type	Max % Strength	Avg Strength	
1804	'ANNUNCIATION'	68	[1]	[S]	100	1	
1805	'CLOSE COMMAND'	68	[1, 1]	[O, O]	50	1	
1806	'EXPLANATION'	68	[1]	[O]	100	1	
1807	'FLOW RATE'	68	[2]	[O]	100	2	
1808	'INSTRUMENTATION'	68	[1]	[O]	100	1	
1809	'MODIFICATIONS'	68	[1]	[S]	100	1	
1810	'OUTAGE'	68	[1]	[O]	100	1	
1811	'OUTLINED'	68	[1]	[S]	100	1	
1812	'PH'	68	[1]	[A]	100	1	
1813	'PLUMBING'	68	[1]	[O]	100	1	
1814	'SHORTEN'	68	[1]	[O]	100	1	
1815	'SIGNAL'	68	[1]	[O]	100	1	
1816	'SIMILAR DAMAGE'	68	[1, 1]	[O, O]	50	1	

The table of matched terms has three additional columns:

- **Matched Sequences:** lists of lengths of sequences matched as a group in the same order as the words composing the term.
- **Match Type:** list of the types of matches in the same order as the words composing the term and in the same order as the matched sequence word groups. The types of matches include:
  - **O** – A word group exactly matches a term in the ontology.
  - **A** – A single-word term matches an acronym in the Aerospace Ontology acronym list.
  - **S** – The stem of a single-word term matches a stemmed word in the Aerospace Ontology.
- **Max % Strength:** An integer indicating how "strong" the match is, expressed as the maximum value of the matched sequences divided by the number of words in the term times 100. In Table 2, the match strength for "close command" is ½ * 100 = 50.

Review of the "Max % Strength" values can focus review of multiword terms on those that are too weak and thus may suggest adding the multiword sequence or some part of it to the ontology. For example, in Table C-3, a review of the terms with 50 percent values might result in adding the phrase "close command" to the Aerospace Ontology, while rejecting the phrase "similar damage."

C.4 Alternative Software for Lexical Analysis

In the course of doing research for this project and others involving lexical analysis of a corpus, an open-source software platform called GATE (i.e., General Architecture for Text Engineering) was found that has a plugin called OpenNLP, which does part-of-speech analysis similar to that performed by SAS®. While obtaining frequency counts for phrases had to be done by additional software written at Johnson Space Center for another project, part-of-speech tagging by OpenNLP was found to scale up well to large corpora as long as the individual text records in the corpus were small in size. This is generally the case for the free-form text fields in PRACA and other problem report records. GATE/OpenNLP may be a better tool for use in future work of this nature than SAS®, not only because GATE is open source and SAS® is costly but because GATE was found to be easy to use and to learn.

C.5 Term Matching Procedure

The application is intended to assist in extending the Aerospace Ontology for use in semantic tagging of documents in additional subject matter areas, disciplines, and businesses. The application compares terms in a list created by lexical analysis with terms in the Aerospace Ontology and writes a table of terms that were found to match and a table of terms for which no match was found in the Aerospace Ontology.

The application is implemented in the Python file "onto_comp.py". The matching procedure is executed by the function: check_terms().

The check_terms function takes several optional arguments:

- *ontopath* – the full path to the ontology's Extensible Markup Language (XML) file.
- *termpath* – the full path to the tab-separated value text file of terms extracted by the data-mining tool.
- *from_raw* – creating a new file of filtered terms with duplicate terms and unused columns in the data-mining table removed.
- *redo*– when true, rerunning the filtered version of the data-mining table.

When both *redo* and *from_raw* are false (the default), the input file of data-mining terms to match is the output file from the last time the matching procedure was executed rather than the original file of data-mining terms. The names of all such files have the form: **Unmatched_terms-***n***.txt**, where *n* is the iteration number. When the input file is **Unmatched_terms-***n*, the output file will be **Unmatched_terms-***n*+1.txt.

Every **Unmatched_terms** table has a "Reject" column on the far right that, if filled in with anything by the ontology developer, indicates that the term has been considered and rejected for inclusion in the ontology on the next iteration.

The columns in the SAS® extract table used by the application are as follows:

- **Term:** The word or series of words extracted from the analyzed set of documents (i.e., database records).
- **Role:** The kind of term extracted, which may be a part of speech, or "Num" for a number or "Prop" for a proper noun. The Role entry is used to filter out numbers and proper nouns from the extract file before attempting to match terms to the ontology.
- **Freq:** The number of occurrences of the term in the set of database records.

The three steps in the procedure are described in detail below.

Step 1: Load and process ontology information from an XML file to create the following three lists:

- Maptext terms – Terms collected from the XML "maptext" of all Aerospace Ontology concepts. The association between terms and concepts is not retained in this data structure.
- Stemmed maptext words – A freeware word-stemming module for Python was used to create stems of the right-most word in each ontology maptext term. The module was downloaded from https://pypi.python.org/pypi/stemming/1.0.1. The "Porter2" algorithm in this module was used to do the stemming. It was chosen because it was the module recommended in the Python.org documentation. The same module provides three other algorithms: Porter, Paice_Husk, and Lovins, some of which are said to be more "aggressive," such as one that stems the verb "succeeded" to the noun "success." Porter2 stems the same verb to the present tense "succeed." However, Porter2 stems are not necessarily (and usually are not) verb infinitives or singular nouns. For example, Porter2 stems both of the words "activate" and "activity" to "activ."
- Abbreviations – The XML ontology file contains a list of acronyms used by STAT. This is also used to match abbreviations found in the term list produced by the data-mining tool.

The maptext terms and abbreviations are converted to all uppercase. The stemmed words are converted to all lowercase because the stemming algorithms expect words to be lowercase.

Step 2: Filter terms in the tab-separated value (TSV) file version of the table produced by the data-mining tool.

If the original TSV file is named *items*.txt, then the new TSV file will be named *items*_filtered.txt. Characters, words, and entire terms are filtered out of the items according to the following rules:

1. Characters are removed from term words.
    - If the character is non-alphabetical (e.g., numbers and punctuation, %). After the removal, a word that initially contained non-alphabetical characters is split into a sequence of (shorter) words separated by a sequence of non-alphabetical characters.
2. Words are removed from terms if the word consists of only one character after Rule 1 is applied.
3. Terms excluded completely:
    - Terms that are null after Rule 2 is applied.
    - Non-printing ASCII characters (e.g., NULL, DLT).
    - Terms with any non-ASCII Unicode characters, as in foreign languages (e.g., umlaut).
    - Terms consisting of more than three words for which the frequency count is less than 4, as reported by the data-mining tool. All other combinations of term length and frequency are accepted.
    - Terms of type "Num" (i.e., numbers) and "Prop" (i.e., proper nouns). The SAS® mining tool designates dates, including alphanumeric dates such as "Feb 1" as type Num.
    - Duplicate terms.

An example of filtering:

The term "CABLE - ISL 1F15940-1" is split into the strings "CABLE," "ISL," and "F." Since the last word contains only one letter, it is removed and the final filtered term is "CABLE ISL."

Step 3: Match terms and record the results in two tables.

A term in the table produced by the data-mining tool is considered to match an ontology term according to the following rules:

1. The entire sequence of uppercase words in the data-mining term exactly matches the uppercase version of an ontology maptext term (the strongest match).
2. The stem of the rightmost word *plus* the sequence of words to its left match exactly to an entire ontology term (e.g., "SOLAR ARRAYS" matches the ontology mapping term "SOLAR ARRAY" exactly by stemming the data-mining plural to its singular form).
3. The rightmost word in the term matches a word in either 1) the list of ontology abbreviations or 2) a word in the list of stemmed ontology words, and the remaining words in the left-hand part of the term sequence match the ontology by either Condition 1 or Condition 2.
4. The word is the last word in the original data-mining term sequence and matches exactly a word in the Removable Words (i.e., stop words) list. Examples of stop words are "some" and "fourth."
5. If the original term consists of only one stop word, it is ignored and not written to either the table of matched terms or the table of unmatched terms.

These rules are applied recursively to multiword terms. Rule 1 is always applied first, since a match for an entire sequence of words in the ontology is a better match than a match for the words taken individually as acronyms or stems.

Output of the Term-matching Application

One table is created to record terms for which no match was found in the ontology, and a second table is created to record terms for which a match was found in the ontology. The frequency of each term reported by the data-mining tool is retained in both tables. The tables are output as TSV files.

Unmatched_terms-*n*.txt

For multiword terms, it is sometimes useful to know whether the single-word matches were partially successful. Two additional "Fail Word" columns were added to the table output. The matching algorithm employed a "greedy" method that attempts to match a phrase from both the left-most word and the right-most word, beginning with an attempt to match the entire phrase. If the algorithm fails to find an exact match, then it searches for the maximal sub-phrase sequence and records the first word failing to match any maptext word in the ontology as the fail word. The first word failing to match in the search starting from the right is recorded in "Fail Word RL" column, and the word failing to match in the search starting from the left is written to the "Fail Word LR" column if different from the "Fail Word RL." The Fail Word could be the problematic word in the sequence that needs to be addressed in the updated ontology.

Matched_terms.txt

The table of matched terms has three additional columns:

- **Matched Sequences** – a list consisting of the lengths of sequences matched as a group in the same order as the words composing the term. The sequence [2, 1, 1] indicates that the first two words in the data-mining term comprised a term in the ontology and the last two words were found as individual entries in either the ontology, the abbreviation list, or the stem list.
- **Match Type** – a list of the types of matches in the same order as the words composing the term and in the same order as the matched sequence word groups. The type matches are:
    - **O** – a word group was matched exactly by a term in the ontology.
    - **A** – a word group consisting of a single word was matched in the list of acronyms used with the ontology.
    - **S** – the stem of the word in a one-word sequence was matched in the list of stemmed words in the ontology.
    - **X** – the word in a one-word group matched a word in the removable words list.
- **Maximum % Match Strength** – This is an integer indicating how "strong" the match is, expressed as the maximum value of the matched sequences divided by the number of words in the term times 100. Examples: a five-word term with a match-word sequence of [1, 3, 1] has a match strength of 100 * 3/5 = 60, and a five-word term with a match-word sequence [5] would have a match strength of 100 (the maximum).

The matching algorithm ensures that the strongest possible match will be found. For example, there might be three different ways to match a given five-word term in the ontology such as:

[1, 1, 1, 2], [1, 2, 2], and [2, 3]

The algorithm will return the [2, 3] match as the "strongest" match because finding two multiword matches in the ontology to subsequences is a better match than the other two matches involving matches to single words. The best match is the one that has the maximum average number of words per group (i.e., the number of words in the term divided by the number of subsequences found in the ontology). A [2, 3] match, therefore, has an average strength of 5 / 2 = 2.5, while a [1, 2, 2] match has a lower average strength of 5 / 3 = 1.67. The [1, 1, 1, 2] match has a strength of 5 / 5 = 1.25, which is the lowest of the three. A match of an entire five-word term to a five-word term in the ontology would, of course, be the strongest with an average strength of 5.

The list of matched sequences and match type entries are related by their positions in the respective lists. The *n*th symbol in the matched type list represents the type of match for the *n*th word group in the Match Sequences entry.

If a matched sequence is "2," its match type must be from the ontology proper because multiword sequences can only be ontology terms. Single matches may be either "O" or "A" for an Acronym match, "S" for the match of the stem of the data-mining term to the stem of an ontology term, or "X" for a match to a stop word. A [2, 1, 1] sequence could have a match type such as [O, O, O], [O, S, S], or [O, A, S], or [O, S, A], etc. The first group is "2" and so could only have an "O" match, while the single-word groups could be "O," "A," or "S" matches.

Step 4: Review Smaller Set of Terms

The ontology developer reviews the list of terms in the last **Unmatched_terms** and **Matched_terms** files and makes additions to the ontology based on the contents of those files, marking the "Reject" column of terms considered but not included in the updated ontology. The developer considers the "Match Strength" values in the **Matched_terms** table to help spot matches of multiword terms that are too weak and, thus, may suggest addition to the ontology of the full term or a multiword sequence portion of the term to the ontology.

Step 5: Iterate

The updated ontology is output as an XML file, and Step 3 is redone using the **Unmatched_terms** file with the "Reject" column marked as needed for the unmatched terms. The **Unmatched_terms** file will be smaller on the next iteration if unmatched data-mining terms have been added as maptext terms to the updated ontology or if any terms have been marked as rejected in the Unmatched_terms file during this iteration.

Appendix D. Basic Process for Customizing and Updating the Aerospace Ontology

D.1 Identify Candidate New Terms or Classes to Add to the Ontology

The purpose of customizing the ontology is to make it useful for indexing and search, based on terms in text fields in a problem report.

Ontology updates can be needed for various reasons:

- New terms (words or phrases) are identified from a new database or set of reports that will be indexed with concept tags. These terms may come from a corpus analysis of the text from the new source to identify the most frequent unique terms. This set of terms can be automatically narrowed down to a spreadsheet of terms that are not matched in the ontology, and the most frequent can be selected as candidate new terms.
- Searching or browsing for the term misses important cases—this could be due to misspelling of words in the text or missing terms in a concept class.
- Concept class content is missing key synonyms or acronyms, or terms seem out of place in a class or there is a missing relationship between terms.
- Concept class seems too broad to narrow down to the correct indexing tag in searches.
- Concept class seems to be in the wrong part of the ontology class hierarchy.

Keep a spreadsheet to record the terms and concept classes that are candidates to add to the ontology.

- Use a Microsoft® Excel® file format for automatic ontology additions. This spreadsheet can be edited during review of the possible addition.
  - At this stage in the process, use the headers shown in Figure D-1 on Sheet 1. Use one row for each candidate change. The headers in row 1 can be assigned in any order using no more than one of each, but as many Member headers as needed.
  - To record a candidate member term, fill in the member field (i.e., the word or phrase, with spaces replaced by underscores).
  - To consider a new candidate concept class, fill in the Subclass field.
  - Use the Comment field to describe the problem or potential solution.

	A	B	C	D	E	F
1	Class	Subclass	Contributor	Comment	Member	Member
2	Decouple	Undock	J. Smith	New member. Not in AO, but "unberth" is in Undock class.	deberth	

Figure D-1. Excel® File Format in Header for Automatic Ontology Additions

D.2 Browse and Search the Ontology for a Missing Term

Determine if the term is in the Ontology version that should be updated. Set the Protégé display to the Entities tab view and type the term into the Protégé search tool in the upper right-hand corner of the display, as shown in Figure D-2. Double-click on the closest term found in the search.

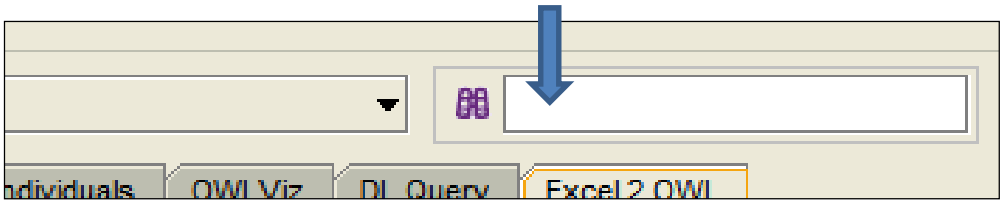


Figure D-2. Search Field in Upper Right Corner of Protégé Display

Note that automatic search can be part of a corpus analysis process. The resulting Excel® file of nonmatching terms would then be manually reduced to a priority set of additions.

If the term is not in the ontology, look for a class that is a potential indexing concept for the term, by browsing the Ontology class hierarchy and using the search tool. Determine appropriate location(s) for the term. It can help to investigate meanings of the term in dictionaries and other sources of definitions.

EXAMPLE: "Deberthing" is not in the ontology. A text context (maintain adequate structural integrity of the MBM-2 during berthing/deberthing of PMA-2 to/from the MBM-2 on the Z1 truss) indicates it is the opposite or reverse of berthing. A search for "berth" and further browsing finds the Undock concept, with members undock and unberth, as shown in Figure D-3. "Deberth" can be added as a member of the Undock concept. Automate stemming of "deberthing" in the text would match it with "deberth." A quick Internet search of dictionaries and thesauruses is a possible follow-up. Indeed, dock (a vessel) is used to define berth. Another synonym, "moor" (i.e., securing a vessel with lines or anchors), is found in the Internet search. If appropriate, this also could be added to the Dock concept class.

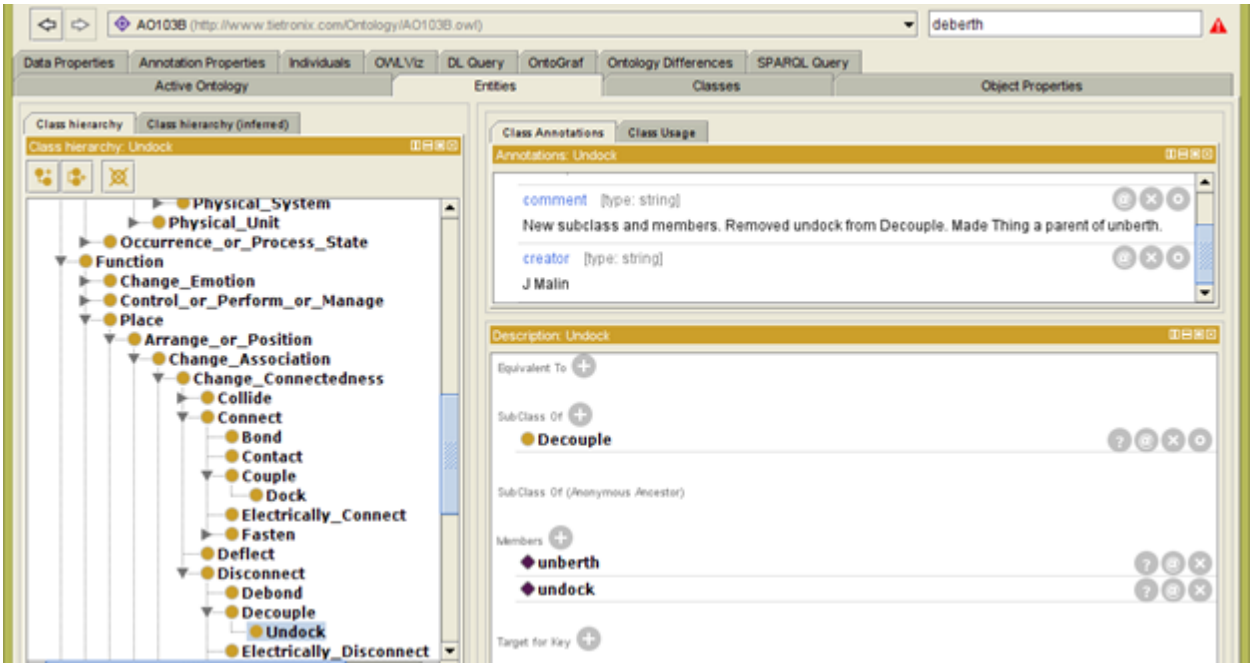


Figure D-3. "Berth" Search and Browse Leads to Undock Concept Class, where "Deberth" Is not Included

If the term is in one or more concept classes in the ontology, check whether one of the identified concept classes correctly reflects the sense of the term in the text where it is found. Do this by comparing the term with the other terms that are members of that class. If the fit does not seem good enough for the needed indexing and search, there may be a missing concept class. Browse and search in the ontology to find potential fits for the term or places where the class would fit. Add to the spreadsheet row the parent Class and candidate name for its new Subclass.

EXAMPLE: "CETA" is a member of the class Acronym, a very general class that would not be a good indexing concept, as shown in Figure D-4. CETA (i.e., crew and equipment translation assembly) will also need to be added to an existing class, Transport_Equipment, or a subclass. A new subclass of Transport_Equipment, for equipment like CETA, could be added.

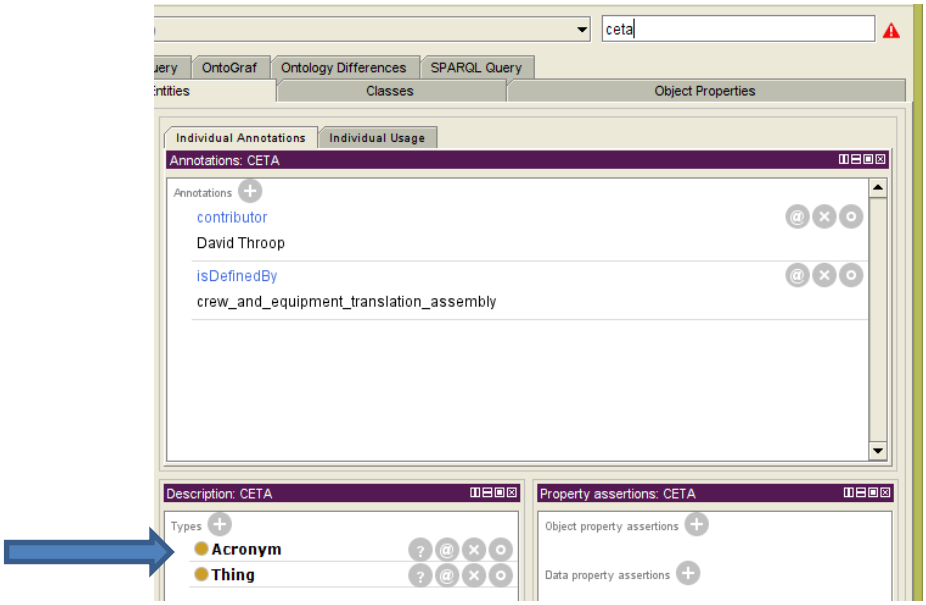


Figure D-4. "CETA" Search Identifies One Class below "Thing," the Universal Parent

EXAMPLE: "SARJ" is a member of the class Joint and the class Acronym. There are only a few joint subsystems in the Joint class, as shown in Figure D-5. It could be split, adding a Joint_Subsystem subclass. Or, even better, the Joint_Subsystem terms could be moved to the Mechanical_Interface class or a new subclass under it. This is better because the grandparent class of Mechanical_Interface is Physical_Structure, while the Joint class parent, Equipment_Part, seems at too low a level for a subsystem such as SARJ.

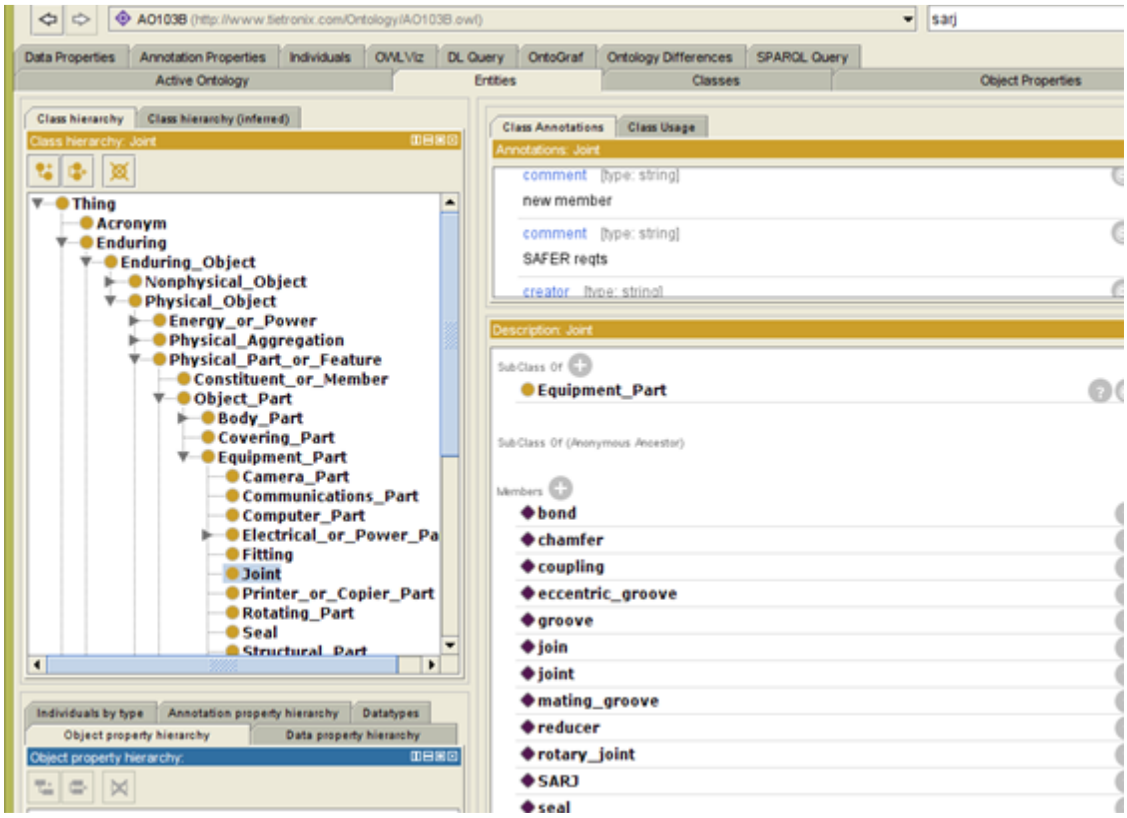


Figure D-5. SARJ Search Leads to Joint Concept Class, with SARJ and Other Members

For these three examples, the resulting Excel® file could be the one in Figure D-6.

	A	B	C	D	E	F
1	Class	Subclass	Contributor	Comment	Member	Member
2	Decouple	Undock	J. Smith	New member. Not in AO, but "unberth" is in Undock class.	deberth	
3	Transport_Equipment?	?	J. Smith	New class for acronym. This is only in the Acronym class. Is crew/equipment transport subsytem. Make this a member of Transport_Equipment or a new subclass of it?	CETA	
4	Mechanical_Interface?	?	J. Smith	New class for acronym. These members are in the Joint class, with parent Equipment_Part. May be better in Mechnical_interface or a new subclass of it.	SARJ	TRRJ

Figure D-6. Excel® File Format for Automatic Ontology Additions

D.3 Add a Class or Member to a Class and Complete the Spreadsheet

Edit the Excel® file to complete the rows of additions to the ontology.

- To add new terms as members of a class, list each new term below a "Member" header.

  - To add a term to multiple classes, add a row for each class.
- To add a new class:
  - Below the "Class" header, enter the name the existing parent of the new class.
  - Below the "Subclass" header, enter the name of the new class.
  - List the members of the new class, each below a "Member" header.
- If the additions will require some manual deletions and class rearrangement, note that in the comment column.
- Complete the file by adding and editing the annotations:
  - Column headers for annotations: Comment, Contributor, Date, Description, Source.
  - Annotations apply to the lowest level class defined in a row. If in a given row, Subclass is empty, all annotations and members will be added to the specified class.

For these examples, the Excel® file could be the one shown in Figure D-7. This file shows that "moor" was chosen to add as a member of the Dock class. It also shows that the possible SARJ/TRRJ class changes were rejected. The definition of TRRJ, "thermal radiator rotary joint," was found to be missing from the Joint class, so it is specified to be a new member. Finally, it shows the addition of the definition of CETA and additional terms to recognize more potential members of the Transport_Equipment class.

	A	B	C	D	E	F	G	H	I	J
1	Class	Subclass	Contributor	Comment	Member	Member	Member	Member	Member	Member
2	Decouple	Undock	J. Smith	New member. Not in AO, but "unberth" is in Undock class.	deberth					
3	Couple	Dock	J. Smith	New member. Closely related to dock and berth.	moor					
4	Placer	Transport_Equipment	J. Smith	New class for acronym. Member is only in the Acronym class. Is crew/equipment transport subsytem. Make this a member of Transport_Equipment or a new subclass of it? Include definition. Generalize with complex expression using System_Unit class (assembly, subsystem, etc.).	CETA	crew_and_equipment_translation_assembly	transporter	transport_equipment	transportation_(System_Unit)	translation_(System_Unit)
5	Equipment_Part	Joint	J. Smith	Include missing TRRJ definition.	thermal_radiator_rotary_joint					

Figure D-7. Filled-out Row for New Class with Members in Ontology Additions Spreadsheet

D.4 Reading and Understanding Complex Expressions

In this example, there are two complex expressions that expand to multiple phrases based on members of the System_Unit class. For example, "transportation_(System_Unit)" would expand to transportation_system, transportation_assembly, transportation_mechanism, and many other

phrases that use terms in the System_Unit class. Here are some rules for understanding complex expressions:

- Lowercase terms represent the words or phrases in that ordered part of the expression.
- Terms that start with an uppercase letter represent classes. They can be in parentheses. The class name in an expression means that every individual from the class will be used in an expansion to multiple member phrases.

  Classes in parentheses ( ) = every individual from the class.

  Classes in brackets [ ] = every individual in the classes and all its subclasses.

D.5 Make Automatic Additions with the Spreadsheet

The Excel® 2 Owl plug-in is used for batch import of ontology additions into Protégé, including classes, members, and annotations.

- Make sure the correct ontology version is loaded, including the Excel® 2 Owl plugin from the Protégé plugin file, and its tab display is visible (i.e., has been activated).
- Perform a Save-As and increment the ontology version number or rename it. The file name format is: AOx.xx.owl, where x.xx is a version number like 1.31.
- Carefully check the Excel® file for misspellings and missing underscores between words in phrases, and save it in .xls format (Excel® 97-2003 format).

  Spaces, #, and % signs are not allowed, and the entries are case sensitive.
- Select the "Excel® 2 OWL" tab in Protégé (see Figure D-8).

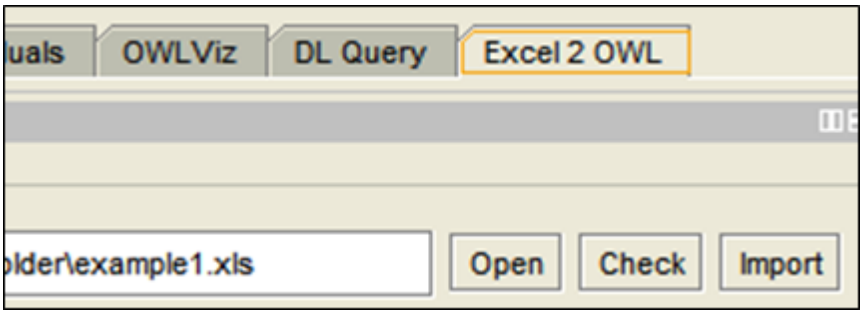


Figure D-8. View of "Excel® 2 OWL" Tab in Protégé and Buttons for Open, Check, and Import

- Click Open to locate the new Excel® file (must be in .xls format).
- Click Check to verify existing and new classes and members (green – class exists; red – new class; blue – new member).
- If needed, click Cancel, investigate existing members or classes by using the Entities view tab, make any corrections that are needed in the file (e.g., misspellings), and start again.
- Click Import to update the ontology (an XLS file named "_classifiers" is also generated but will not be needed for this application).

- To verify new additions, search for new classes or members (use search tool) and review modifications in the Entities Tab view.

D.6 Manual Changes to Remove, Delete, or Rearrange

After making modifications and additions to the Ontology, it may be necessary to manually and interactively remove a moved member from an old class or rearrange the class hierarchy. This can includes moving concepts in the hierarchy up or down a level.

The spreadsheet should include these needed manual changes in the comment column. As each of these changes is accomplished, edit the comment annotation in Protégé so that it no longer says a change is needed but states that a change was made.

For example, after checking the Excel® 2 Owl import, another class, Vehicle, is found with "transporter" as a member. Since that class is a subclass of Transport_Equipment, "transporter" should be deleted from the Vehicle Class. This is noted in the spreadsheet before the file is imported.

In addition, each new member will need to be made a subclass of the universal parent, "Thing," as well as its direct parent concept class.

D.7 Deleting Members from a Class or Adding a Parent Class

Select the Classes Tab and locate the Description pane.

Search for the member or class to change. In this example, a search for "transporter" and selection of the Vehicle class produces the Description pane shown in Figure D-9.

To remove "transporter" from the Vehicle class, click on the X to the far right of the term in the Description Pane. Change the class annotation also.

Then click on "transporter." An Entities Tab view is shown, and in the Description Pane the parent classes (immediate and top-level) are shown under the Types heading. If the Thing class is not one of the parents, it will need to be added. Select the + button in the pane, select the Class Hierarchy tab in the pop-up, and select Thing.

This method assumes that the member that is removed from a class is still in the ontology in another class. To accomplish a complete deletion of a member from the ontology (e.g., to correct an error such as misspelling), bring up the Individuals tab and delete the member from the long list in the Individuals pane.

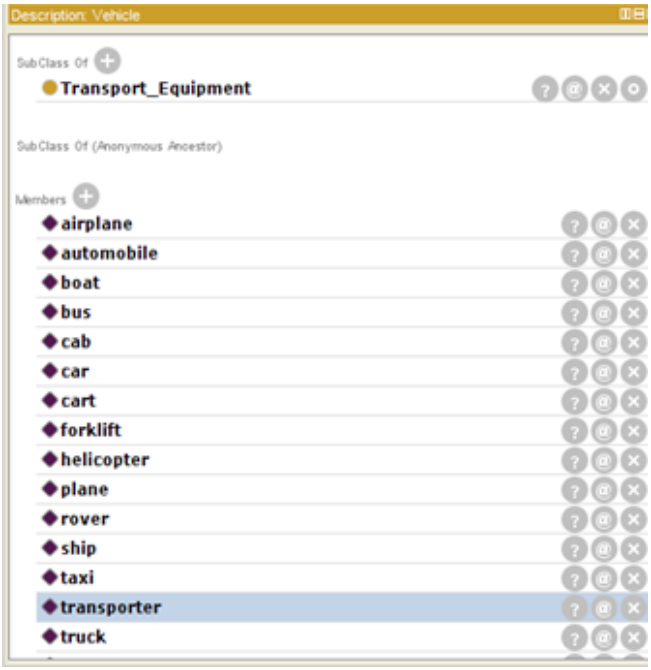


Figure D-9. Description Pane, with "Transporter" Member of Vehicle Class

D.8 Export Ontology to XML

The Export Ontology to XML plug-in is for exporting an ontology to the .xml file that is needed for STAT processing. Select "Export ontology to XML" from the File drop down menu. (Make sure you have loaded this plug-in.)

- Create a file name and file location.
  - The title format for the new version is: Vers x.xx Aerospace Ontology, where the version number (e.g., 1.31) corresponds to the Aerospace Ontology version of the .owl file.
- Click Save to begin export.
- Enter tag names for the main Ontology classes when prompted:
  - Tag name for Acronym: Acronym
  - Tag name for Enduring: Enduring
  - Tag name for Function: Function
  - Tag name for PROBLEM: PROBLEM
  - Tag name for Property_Value: Property_Value
  - Tag name for UserDefinedClassifier: UserDefinedClassifier

Appendix E. Data Visualization

A key enabler of data trend analysis is to have an effective tool for users to query the data and to visualize the output. This assessment used two data query and visualization tools, Tableau® and Flamenco. Tableau®, a commercial off-the-shelf (COTS) tool, was used for its strength as an intuitive, state-of-the-art data visualization tool. Flamenco was used for its strength as on open-source, multifaceted search tool.

E.1 Tableau® Visualization, Version 8.2

Tableau® Desktop and Tableau® Reader are multi-platform, COTS software programs that were procured to assist the NESC team assessment using data visualization. The Tableau® Desktop built the connection to data sources for querying, calculating, code generating, and graph building, to facilitate the construction of data visualizations. The Tableau® Reader (freeware) allowed for viewing, filtering, sorting, exporting, and printing; facilitating the interactive visualization of the files produced by the Tableau® Desktop. Using both Tableau® Desktop and Reader in combination provided the NESC assessment team with the ability to visualize patterns into this large data set and drill down via mouse click.

As the team and discipline experts interacted with the capabilities querying and displaying data, Tableau® Desktop was used to enhance visualization and data search capabilities.

Tableau® Desktop standalone version was utilized to develop the visualization dashboards and produce the workbook files that are used by the Tableau® Reader. The files produced by Tableau® Desktop are a standalone workbook package that contains background images, Excel® files, and data extracts. These files were used to construct worksheets within the Tableau® workbook and facilitated the creation of a mouse point-and-click environment within which relevant areas could be manipulated in order to analyze the data visually. Tableau® Desktop produced the workbooks for the Tableau® Reader to visualize and analyze structured and unstructured data. Tableau® Reader not only displayed structured data visually but also allowed searches of unstructured data and displayed it in a visualization.

The challenges encountered when adding visualization included determining the number of graphs allowed to a single screen while not overwhelming the discipline experts. In managing thousands of records and trying to create a meaningful visualization, basic techniques and tips were used in this assessment:

- Understand the data size and cardinality.
- Determine what the visualization should display.
- Choose the right graph for the data.
- Understand which systems were of interest to a discipline expert.
- Keep it simple.

E.1.1 Typical Example of a Tableau® Dashboard

Figure E-1 is a full view of a Tableau® visualization dashboard that was developed during the assessment. A dashboard is a composition of sheets, and each sheet is a different view of the data. It is similar to the dashboard of a vehicle, with multiple individual gauges and displays, each yielding different information, or different perspectives, of what is happening underneath the hood. A visualization can have one or more dashboards, and each sheet within a dashboard can be viewed individually.

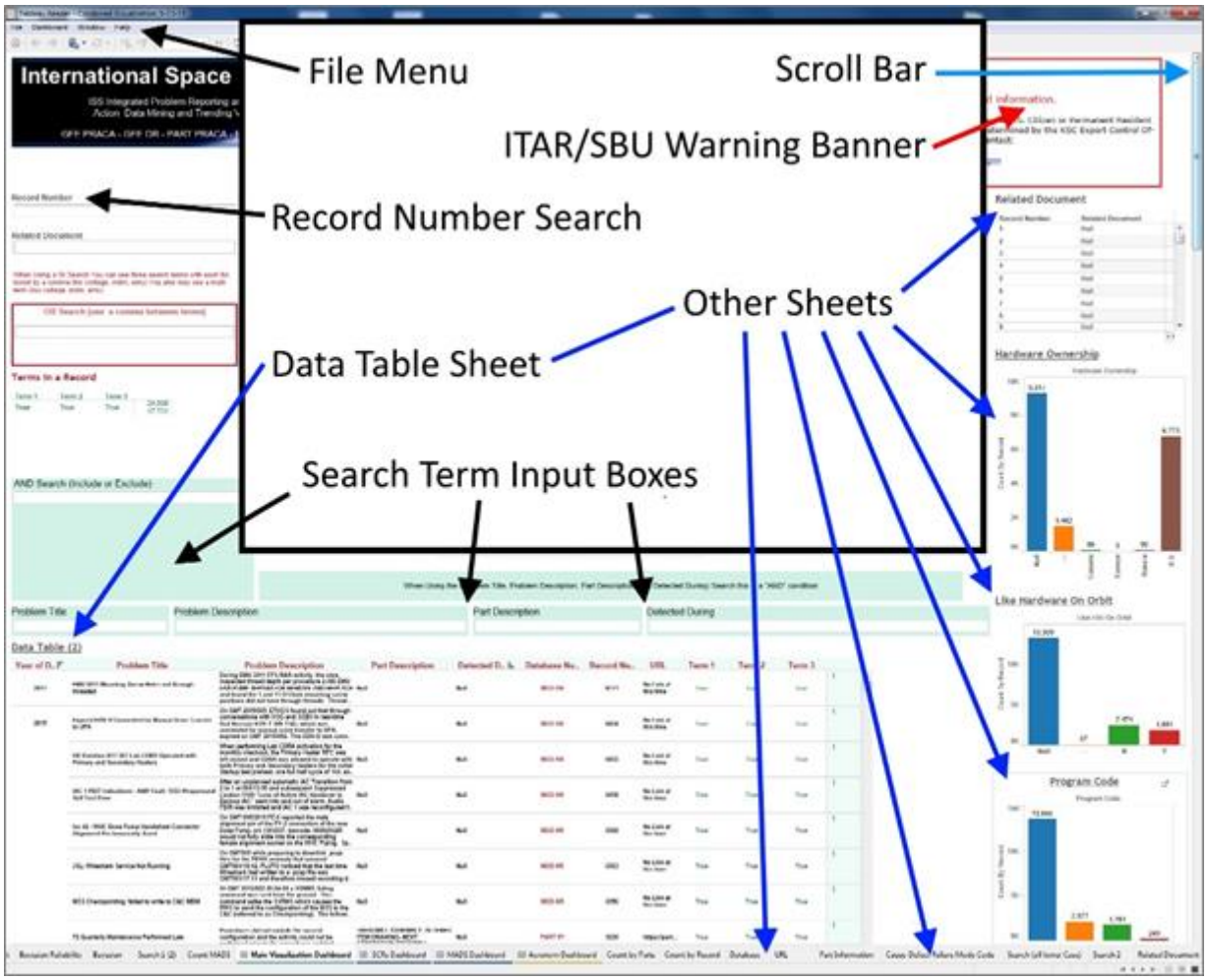


Figure E-1. Tableau® Reader Dashboard

Each sheet within the dashboard, whether a table or graph, has a sheet name descriptive of the data that it visualizes. There is also a control on each sheet to allow access to a full view of only that sheet.

The Hardware Ownership sheet is shown in Figures E-2 and E-3. The sheet title is descriptive of the type of information displayed on the sheet. It shows the number of records containing the associated information. Notice the counts to the left are records with no ownership field (9,317), followed by records with an ownership field with a blank field (1,462). Changing the focus of one or more of the other sheets in the dashboard that are connected to this sheet changes the context and can thereby change these counts. For example, going to the Data Table sheet in Figure E-1 and clicking the year 2012 inside the "Year of Detected Date" column changes the current dashboard context to all data pertinent to the year 2012. Every sheet in the dashboard will then display only data from the year 2012, and the counts in this view will change accordingly.

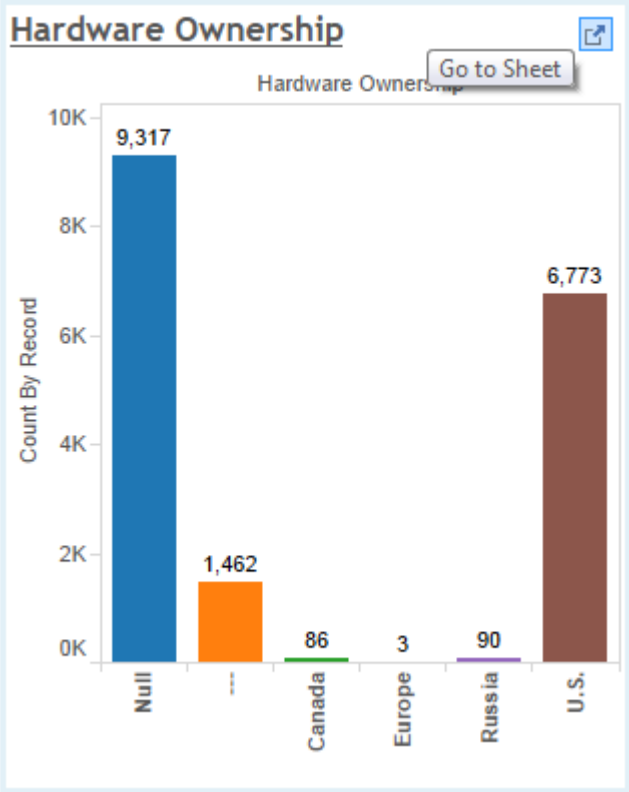


Figure E-2. Sample Hardware Ownership Sheet

To the upper right of every sheet is a button that returns to a full view of the source sheet. The full "Hardware Ownership" sheet is shown in the Tableau® Reader window in Figure E-3.

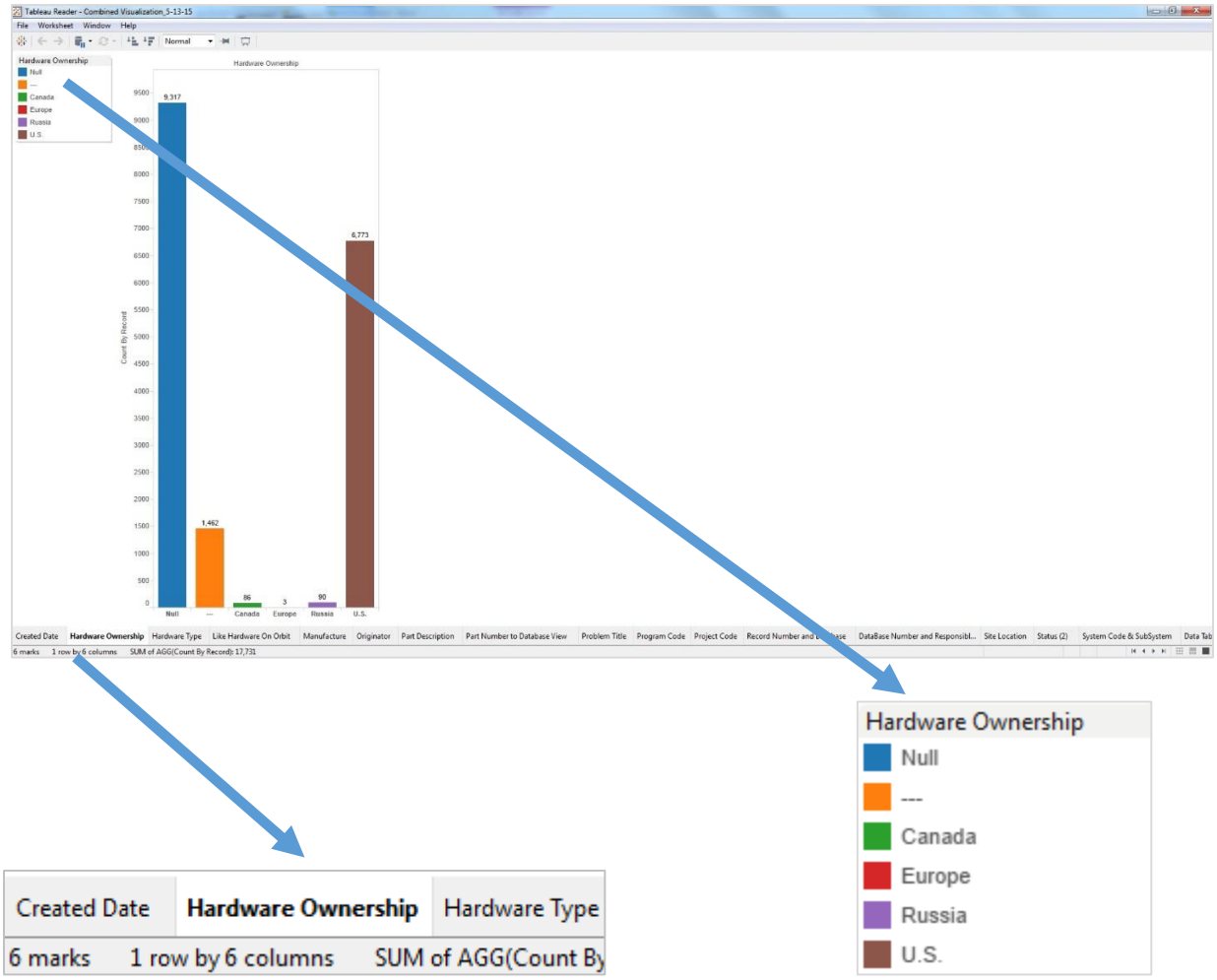


Figure E-3. Hardware Ownership Sheet Detail

Notice the names on the tabs at the bottom of the Tableau® Reader window in Figure E-3. Each tab bears the name of the sheet as it appears in the dashboard. Thus, there are two ways of going from the dashboard to a source sheet:

1. Clicking on the ( ) icon to the upper right of a sheet on the dashboard.
2. Clicking on the tab at the bottom of the screen that bears the name of the desired source sheet. There is also a dashboard tab at the bottom of the window.

The data visualization dashboards are designed to depict the multidimensional aspects and measures of problem reports. The International Space Station (ISS) dashboard shown in Figure E-4 has six zones of interest: one query zone and five display zones.

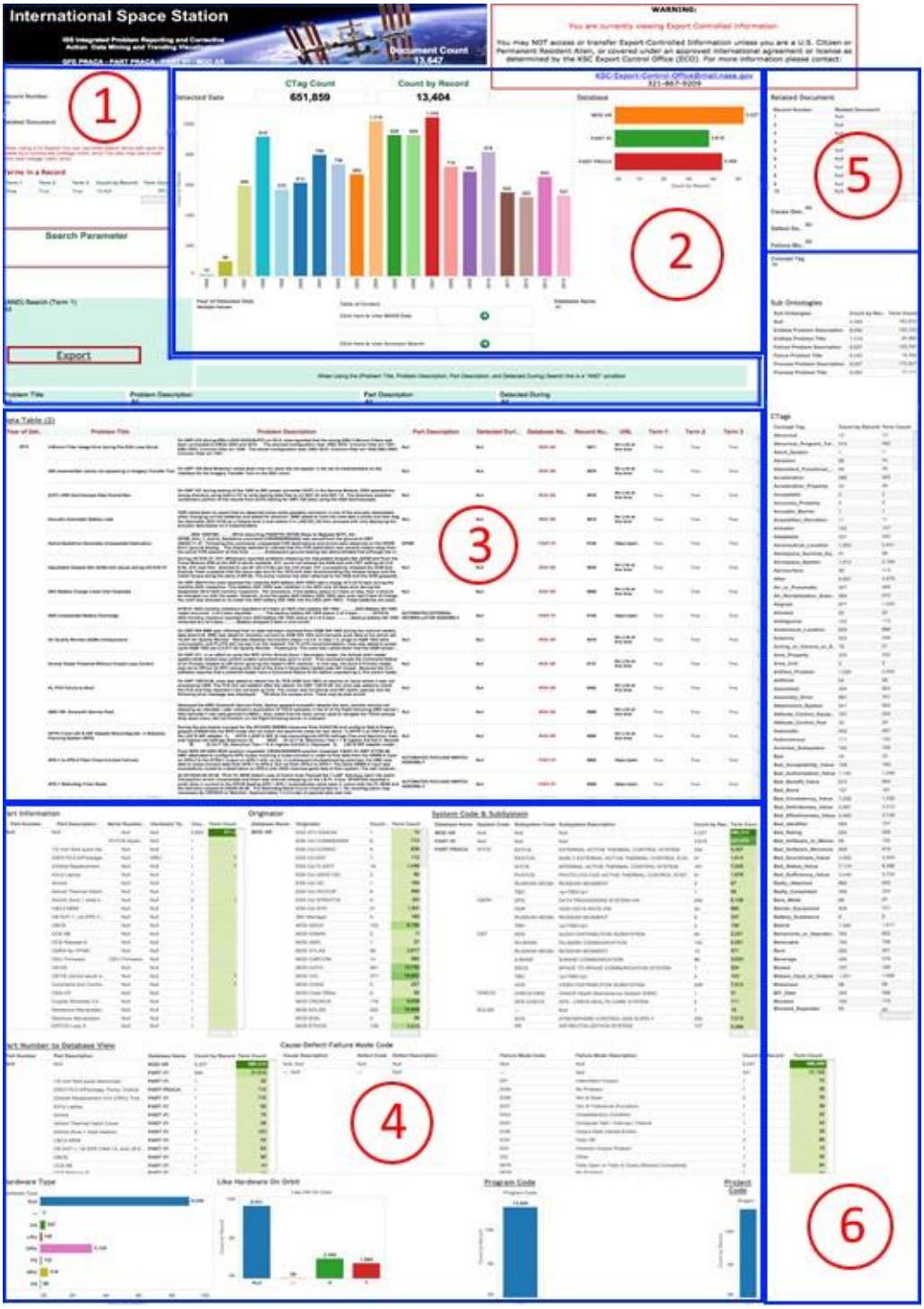


Figure E-4. Data Visualization Dashboard

The numbers in Figure E-4 correspond to the following zones:

- Zone 1 (Figure E-5): text entry area used to query the combined data sets.
- Zone 2 (Figure E-6): record count summary showing occurrences detected per year and total records per database.
- Zone 3 (Figure E-7): records table.
- Zone 4 (Figure E-8): various other important counts, such as a count by part number and a count by cause codes.
- Zone 5 (Figure E-9): records related to the currently selected record, with an ability to filter results by cause, defect, or failure mode.
- Zone 6 (Figure E-10): tables of record counts associated with sub-ontologies and concept tags, the latest update to the visualization, and a text filter. The entered text filters the concept tags table to those tags containing the entered text and subsequently filters all other zones on the dashboard to data pertaining to those concept tags.

Figure E-5. Text Entry Area

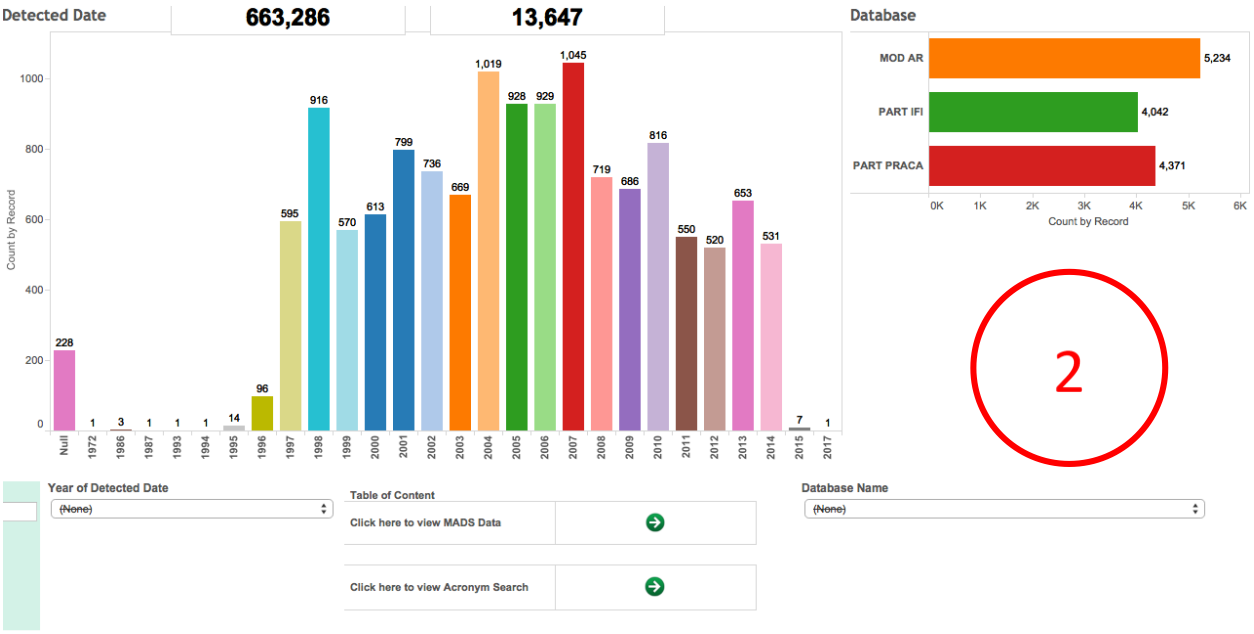


Figure E-6. Record Count Summary

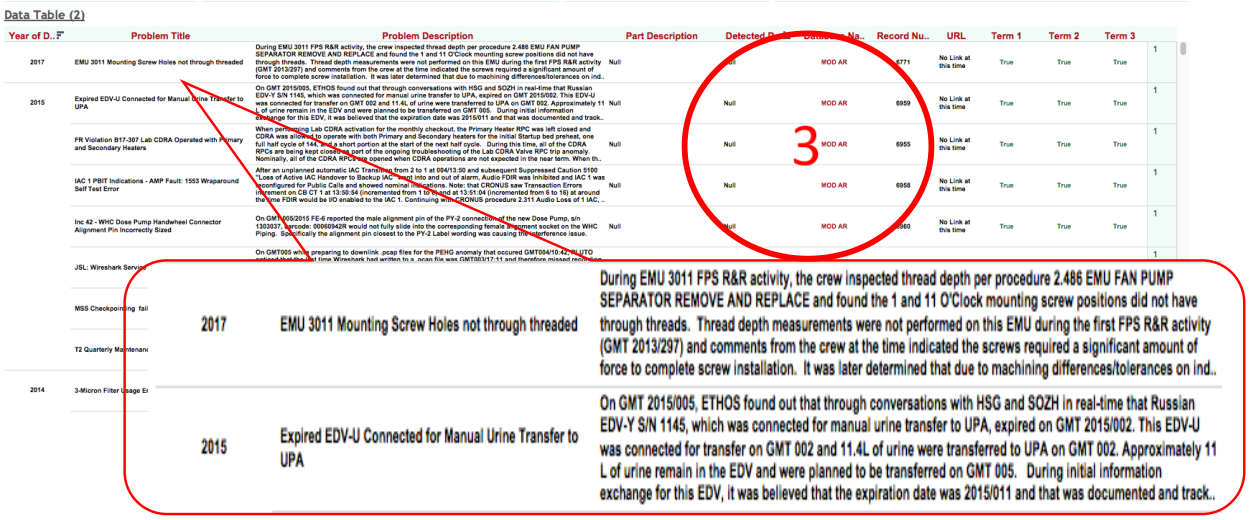


Figure E-7. Records Table

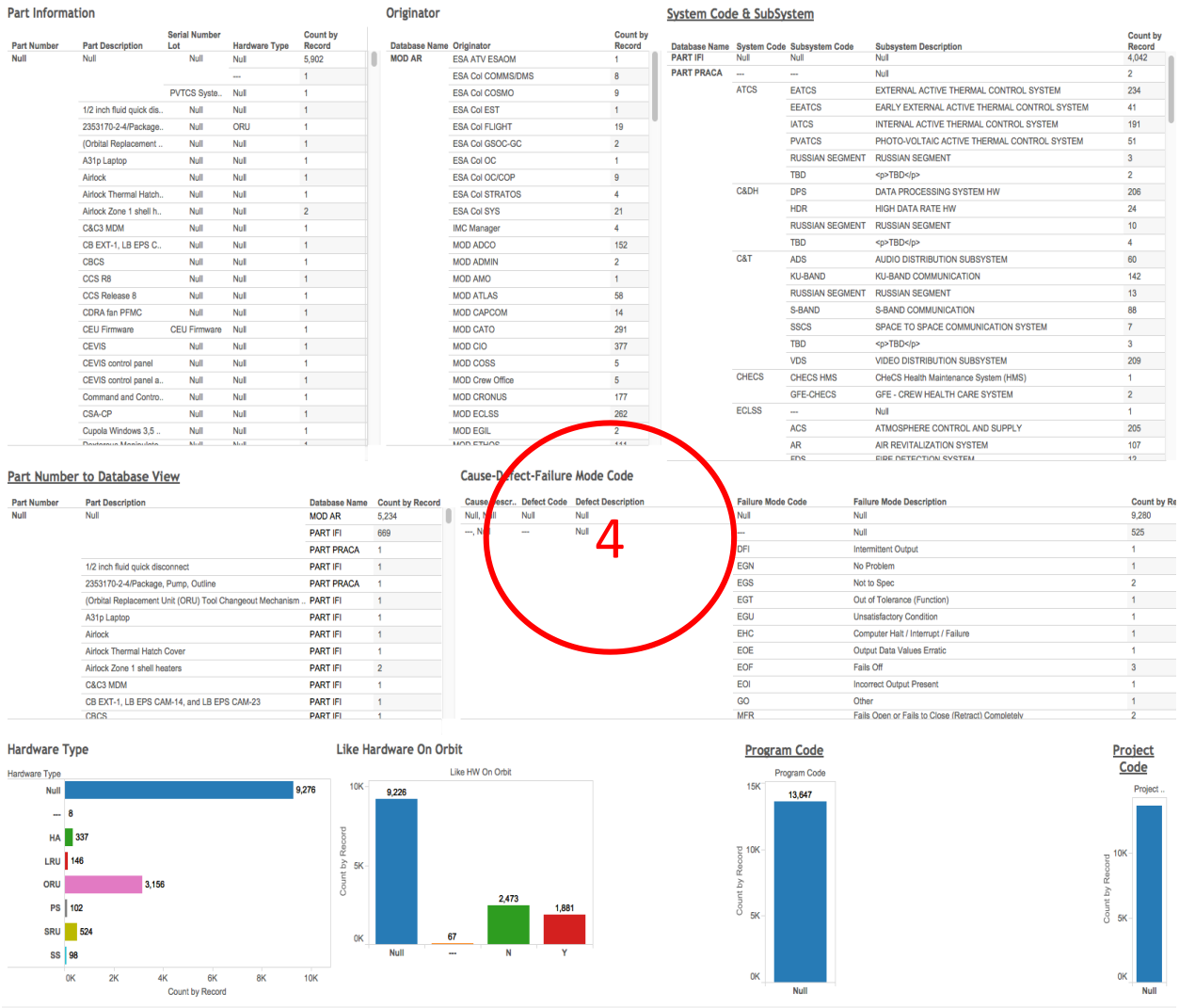


Figure E-8. Other Counts

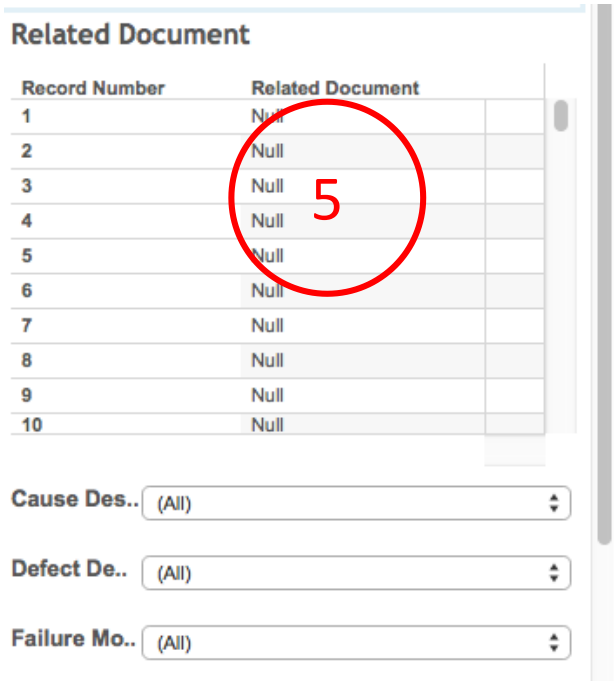


Figure E-9. Records Related to Currently Selected Record

Concept Tag

Sub Ontologies

Sub Ontologies	Count by Record	Count by Terms
Null	4,371	164,340
Entities Problem Description	9,211	136,469
Entities Problem Title	7,506	21,519
Failure Problem Description	9,149	126,445
Faiure Problem Title	6,418	19,889
Process Problem Description	9,229	176,992
Process Problem Title	6,231	17,632

CTags

Concept Tag	Count by Record	Count by Terms
Abnormal	13	13
Abnormal_Program_Ter..	418	497
Abort_System	1	1
Abrasion	59	75
Absorbent_Functional_S..	71	78
Acceleration	291	310
Acceleration_Property	44	46
Acceptable	2	2
Accuracy_Property	2	2
Acoustic_Barrier	1	1
Acquisition_Deviation	11	11
Actuator	123	148
Adaptable	227	246
Aeronautical_Location	1,942	2,487
Aerospace_Survival_Equ..	95	103
Aerospace_System	1,960	2,216
Aerosurface	99	113
After	4,924	5,057
Air_or_Pneumatic	418	481
Air_Revitalization_Subsy..	402	583
Aligned	890	1,020
Allowed	22	22
Ambiguous	105	116
Anatomical_Location	663	696
Antenna	259	354
Arcing_or_Corona_or_St..	19	21
Area_Property	227	234
Area_Unit	3	3
Artifact_Problem	1,959	2,074
Artificial	55	57
Assembler	567	680
Assembly_Error	690	75[illegible]
Attachment_System	5[illegible]	613
Attitude_Control_Equip..	170	248
Attitude_Control_Part	23	30
Automatic	462	506
Autonomous	114	123
Avionics_Subsystem	142	146
Bad	25	25
Bad_Acceptability_Value	14[illegible]	165
Bad_Authorization_Value	1,171	1,277
Bad_Benefit_Value	833	866
Bad_Bond	160	185
Bad_Consistency_Value	1,246	1,347

6

Figure E-10. Tables of Record Counts Associated with Sub-ontologies and Concept Tags

The ISS dashboard displays data from problem reports (i.e., GFE DR, GFE PRACA, PART PRACA, PART IFI, and MOD AR) combined using multiple sheets to add dimensions and measures that allow for drilldown to a single record. Performing a search surveys the combined problem report data set. Capability was added later to search across the MADS and SCR data, and related acronyms were also displayed. A part description like "MDM" could be entered into the combined dashboard Parameter Search field (i.e., the red box in zone 1 of Figure E-5) and MADS, SCRs, and acronyms would all be searched and the data would be displayed (i.e., GFE PRACA, PART PRACA, PART IFI, and MOD AR), combined using multiple sheets to add dimensions and measures that allow for drilldown to a single record. Performing a search surveys the combined problem report data set. Capability was added later to search across the MADS and SCR data, and related acronyms were also displayed. A part description like "MDM" could be entered into the combined dashboard Parameter Search field (i.e., the red box in zone 1 of Figure E-5) and MADS, SCRs, and acronyms would all be searched and the data would be displayed visually.

To enhance visualization during the initial design process, fields were added to the Tableau® dashboard. The two major types of fields added were calculated fields and search fields. These fields were either visible or hidden, depending upon whether it was a query interim step or the final step in the visualization. Only final steps were visible.

The visualization was divided into two areas: categorical data called "dimensions" and quantitative data called "measures."[2] This was where the data roles were separated within the dashboard. Dimensions created an axis of categories and headings, while measures created an axis showing continuous scale. In each case, decisions were made to make the fields discrete or continuous. In most cases, dimensions were discrete and measures were continuous. The Detected Date chart (see Figure E-11) represented a time dimension that is discrete. The Hardware Type chart (see Figure E-12) represented an axis showing continuous scale.

[2] Different *zones* are layout containers, either horizontal or vertical. *Field* is a dimension (field from the database) in a layout container. *Categorical data* is the statistical data type consisting of categorical variables or of data that has been converted to that form, for example, as grouped data.

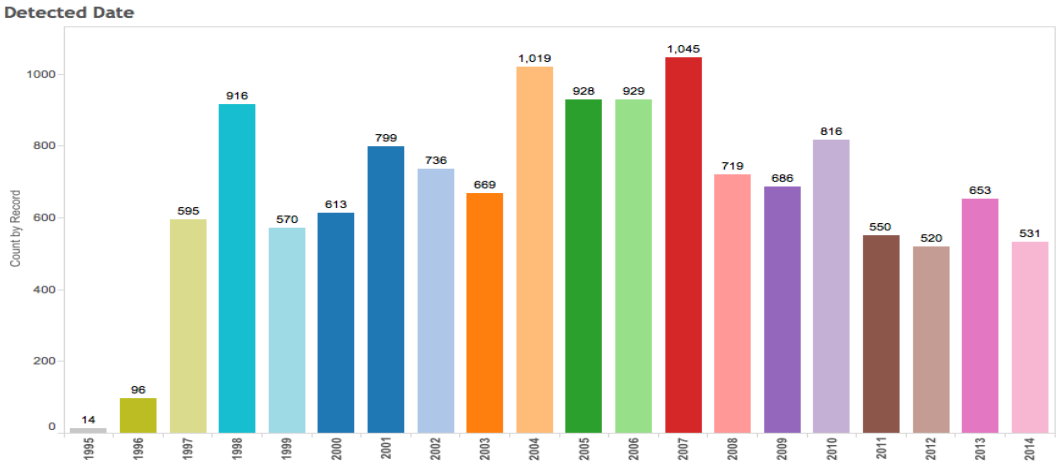


Figure E-11. Detected Date Chart

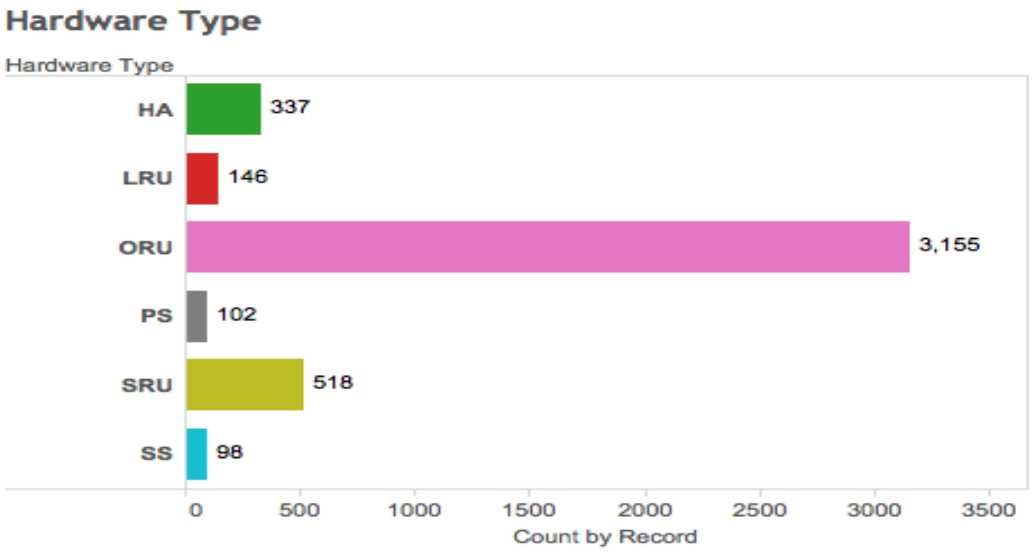


Figure E-12. Hardware Type Chart

A search enhancement was developed and added to the Tableau® combined dashboard. This search enhancement provided the ability to search multiple fields. Further development of the search enhancement allowed not only multiple fields search but multiple dashboards within the Tableau® workbook to be searched at the same time. This provided the ability to search up to three terms (i.e., mdm, software, rpc) in a search field (see Table E-1). The table below indicates how many records contained a single term, a combination of two terms, or all three terms. *True* in every column indicates that all three terms were found in the number of records shown in the Total. *True* in only one column, with *False* in the others, for some rows, indicates that for the total records shown in that row, only one term of the three will be found in those records. Single words and phrases (i.e., rpcm, critical data, last command) were also used within the search (see

Table E-2). In query language, the statement is an "or" rpcm or "critical data" or "last command" and displays the true or false counts for each record searched.

Table E-1. Search Terms

mdm, software, rpc			
Term 1	**Term 2**	**Term 3**	**Total**
False	False	False	21,742
		True	775
	True	False	1,440
		True	39
True	False	False	625
		True	118
	True	False	135
		True	**34**

Table E-2. Multi-search Terms

rpcm, critical data, last command			
Term 1	**Term 2**	**Term 3**	**Total**
False	False	False	24,151
		True	45
	True	False	2
True	False	False	666
		True	**41**
	True	False	3

E.1.2 Tableau® Search and URL Code

Because data input from numerous users to the different source databases was not consistent in alphabetic case used, further modification was made to the Tableau® Desktop code to make searches case insensitive. Since the off-the-shelf Tableau® search query was case sensitive when using a parameter search (i.e., search using the text entry area), a hidden calculated field was used to change the text case of the data to be searched into lowercase to normalize all the data. The case format of the data source records was retained and was used when displaying record data. Changing the data to be searched into lowercase was the best way to accomplish a normalized query (see Figure E-13).

Search 1

```
IIF(ISNULL([Problem Description]),"",LOWER([Problem Description]))+
IIF(ISNULL([Problem Title]),"",LOWER([Problem Title]))+
IIF(ISNULL([Detected During]),"",LOWER([Detected During]))+
IIF(ISNULL([Part Description]),"",LOWER([Part Description]))+
IIF(ISNULL([Part Number]),"",LOWER([Part Number]))+
IIF(ISNULL([Record Number]),"",LOWER([Record Number]))+
IIF(ISNULL([Cause Description]),"",LOWER([Cause Description]))+
IIF(ISNULL([Defect Description]),"",LOWER([Defect Description]))+
IIF(ISNULL([Failure Mode Description]),"",LOWER([Failure Mode Description]))+
IIF(ISNULL([Subsystem Description]),"",LOWER([Subsystem Description]))
```

Term 1

```
CONTAINS([Search 1], TRIM([Search Parameter])) OR CONTAINS([Search 1],TRIM(LEFT([Search Parameter],FIND([Search Parameter],
",")-1)))
```

Term 2

```
(], ","))), ",") = False THEN
(IF IF CONTAINS(TRIM(RIGHT([Search Parameter],LEN([Search Parameter])- FIND([Search Parameter CONTAINS(lower([Search 1]),
lower(TRIM(RIGHT([Search Parameter],LEN([Search Parameter])- FIND([Search Parameter], ","))))) THEN 1 ELSE 0 END)
ELSEIF CONTAINS(lower([Search 1]), lower(LEFT(TRIM(RIGHT([Search Parameter],LEN([Search Parameter])- FIND([Search Parameter],
",")))
,FIND(TRIM(RIGHT([Search Parameter],LEN([Search Parameter])- FIND([Search Parameter], ","))), ",")-1))) THEN 1 ELSE 0
END) = 1
```

Term 3

```
CONTAINS([Search 1],TRIM(RIGHT([Search Parameter], LEN(RIGHT([Search Parameter], LEN([Search Parameter]) - FIND([Search
Parameter], ","")))
- FIND(RIGHT([Search Parameter], LEN([Search Parameter]) - FIND([Search Parameter], ",")), ","))))
```

String Match

```
[Term 1] or [Term 2] or [Term 3]
```

Figure E-13. Search Sample Code

Uniform Resource Locators (URLs) were added to the visualization dashboards workbook. This enabled users with proper permissions to link directly back to the source record of a particular anomaly. A "URL Action" making use of hidden calculated fields was programmed to construct the URL based on the data source, record number, and how the destination web server processed the URL. The original data source hyperlinks used different suffixes and prefixes to retrieve the records. Thus, the challenge was to use the correct suffix and prefix for each data source URL hyperlink and implement it with a standalone reader. Another challenge was searching identical document numbers from different data sources (see Figure E-14).

```
//URL with suffix
RIGHT( [URL], LEN([URL])-FIND([URL],"<Record>")-6)

//URL with Prefix based on data source
IF [Database Name] = "GFE PRACA"
THEN "https://qfed-sma.jsc.nasa.gov/PRACA/Common/Common.aspx?DocumentNumber="
ELSEIF [Database Name] = "GFE DR" THEN "https://qfed-
sma.jsc.nasa.gov/QARC/Pages/Report01.aspx?ID=0&Results=On&DocumentNumberField="
ELSEIF [Database Name] = "PART PRACA" THEN "https://part.iss.nasa.gov/show_bug.cgi?id="
ELSEIF [Database Name] = "PART IFI" THEN "https://part.iss.nasa.gov/show_bug.cgi?id="
END
//URL with suffix based on data source
IF [Database Name] = "GFE PRACA"
THEN ""
ELSEIF [Database Name] = "GFE DR" THEN "&"
ELSEIF [Database Name] = "PART PRACA" THEN "&ctype=pdf"
ELSEIF [Database Name] = "PART IFI" THEN "&ctype=pdf"
END
```

Figure E-14. URL Sample Code

E.1.3 Supporting Dashboards

Supporting dashboards were added to enhance the visualization experience by giving more information. Three dashboards supported the combined problem reporting visualization.

- Acronym (see Figure E-15)
- MADS (see Figure E-16)
- SCRs (see Figure E-17)

Each supporting dashboard was designed to give access to each of the three data sources (i.e., Acronym, MADS, and SCRs) without being connected to the original data source, but providing a link to each original data source if required.

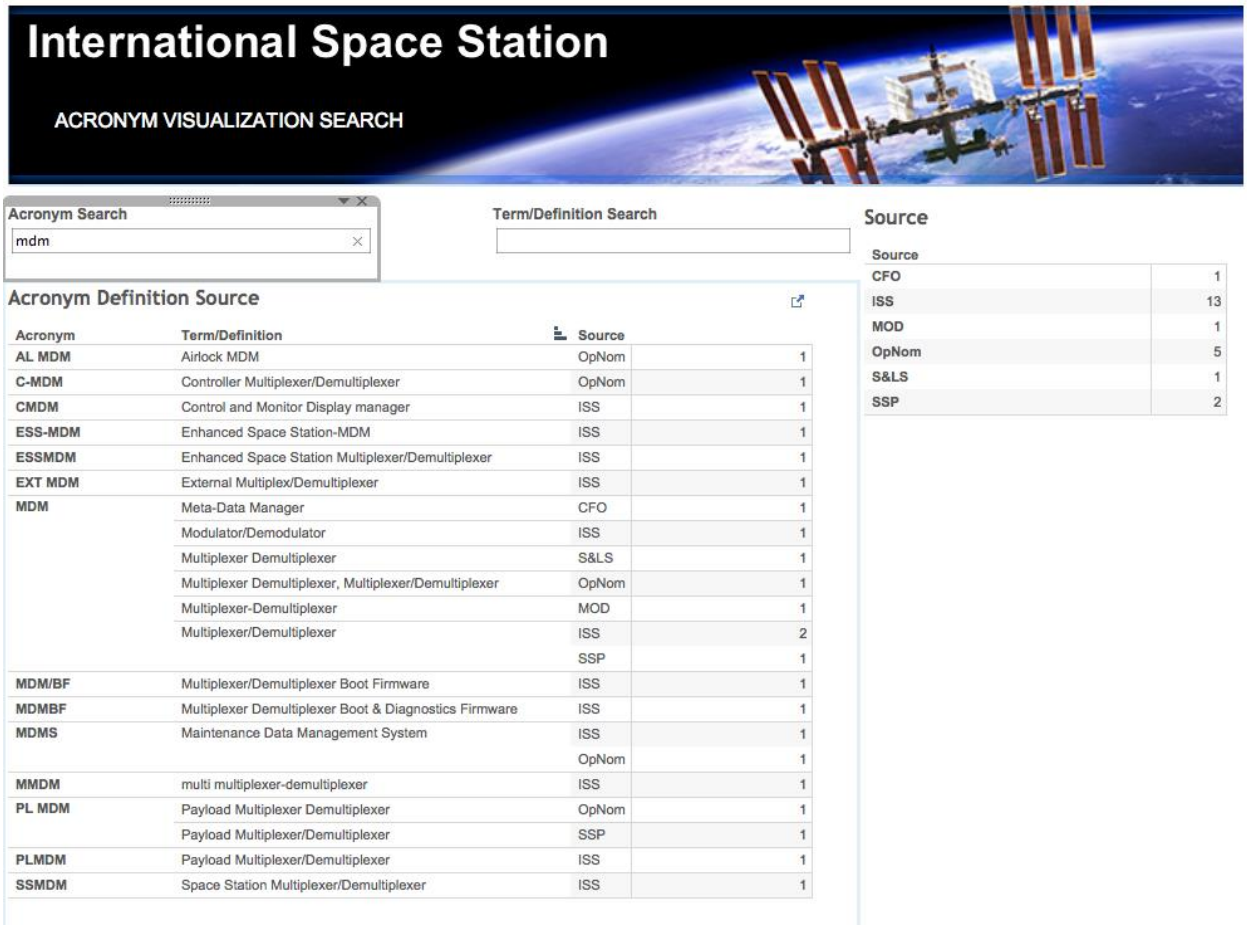


Figure E-15. Acronym Dashboard (data source: http://www6.jsc.nasa.gov/AcronymCentral/scripts/index.cfm)

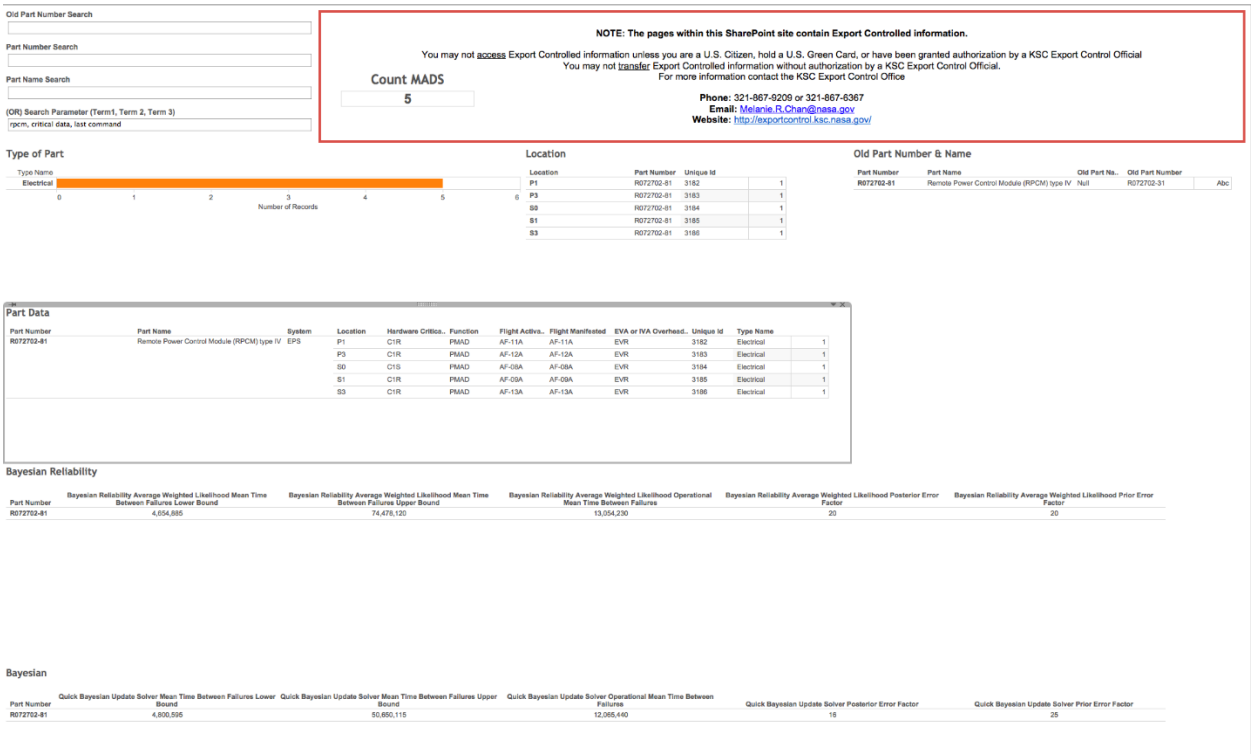


Figure E-16. MADS Dashboard (data source: https://iss-www.jsc.nasa.gov/madsx/f?p=mads:1:0:::::)

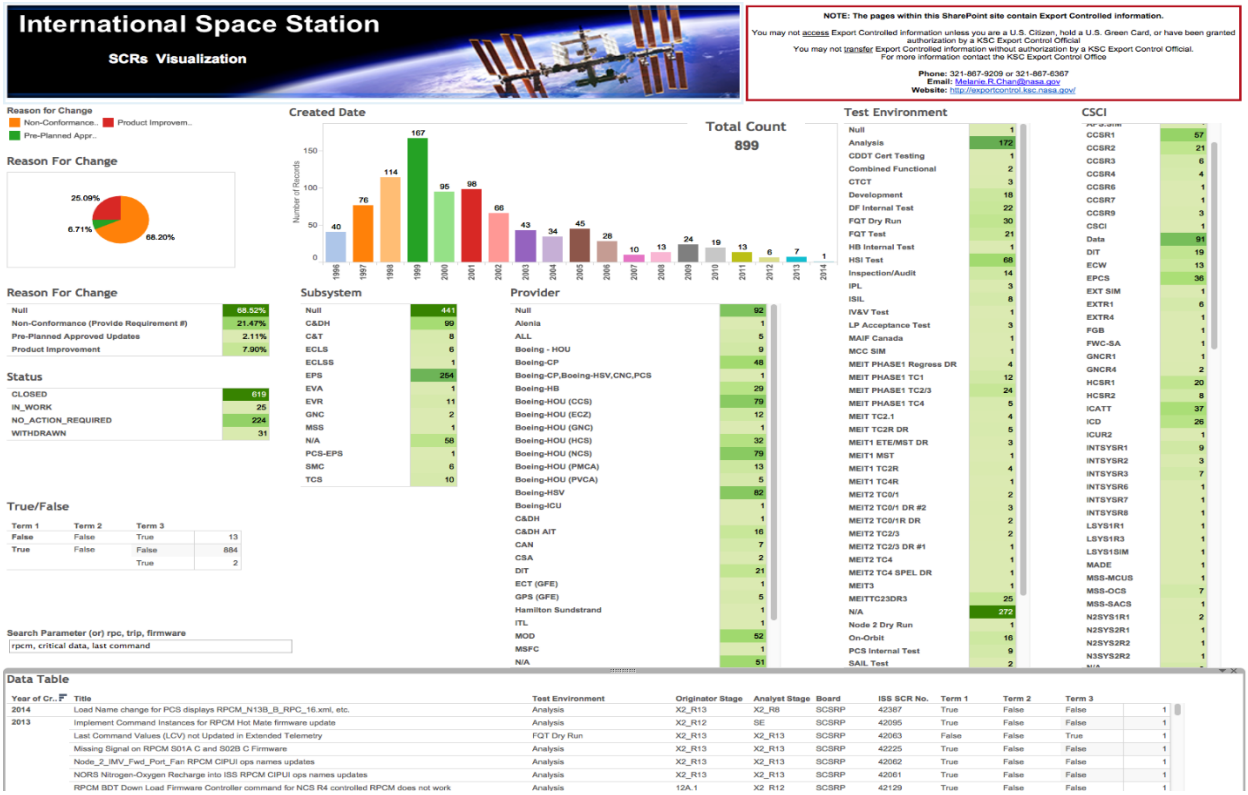


Figure E-17. SCR Dashboard (data source: https://pvcsweb.jsc.nasa.gov/external_access/browse/browse_filter.cgi)

E.2 Flamenco

This is an illustrated scenario showing how the Flamenco visualization containing concept tags can be used to find information in anomaly reports related to an issue of interest to a domain expert.

First, select the Flamenco database that has GFE PRACA data (see Figure E-18).

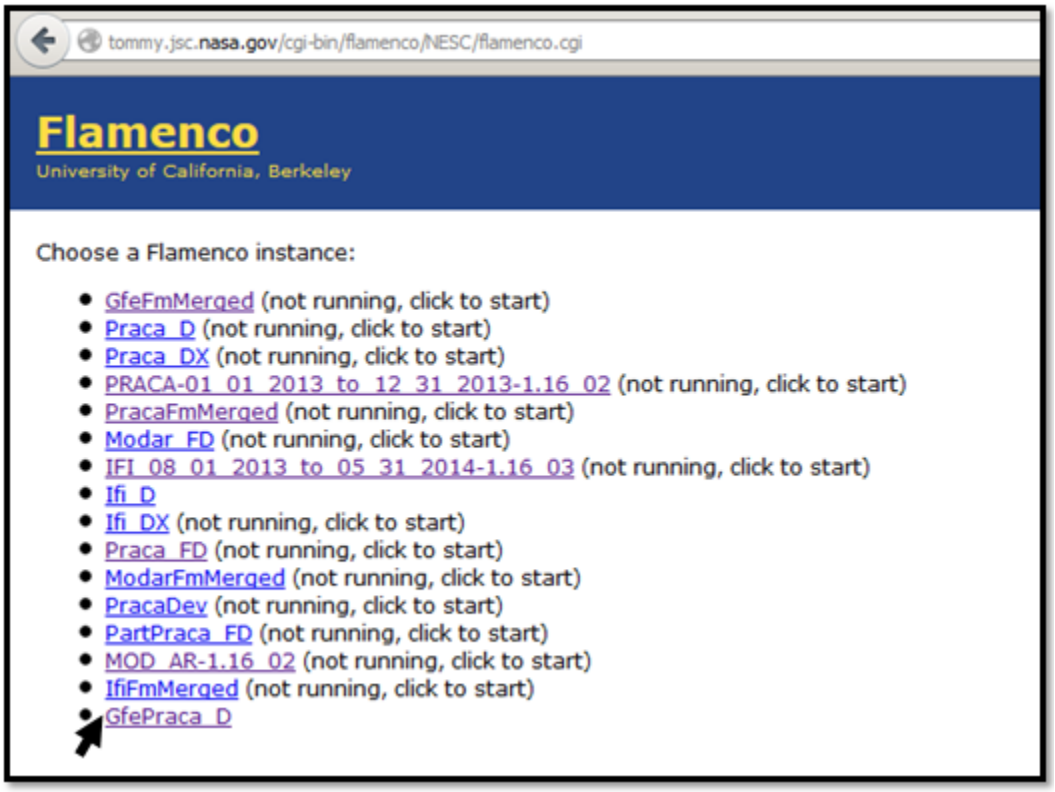


Figure E-18. List of Available Flamenco Databases

Figure E-19 shows the resulting view.

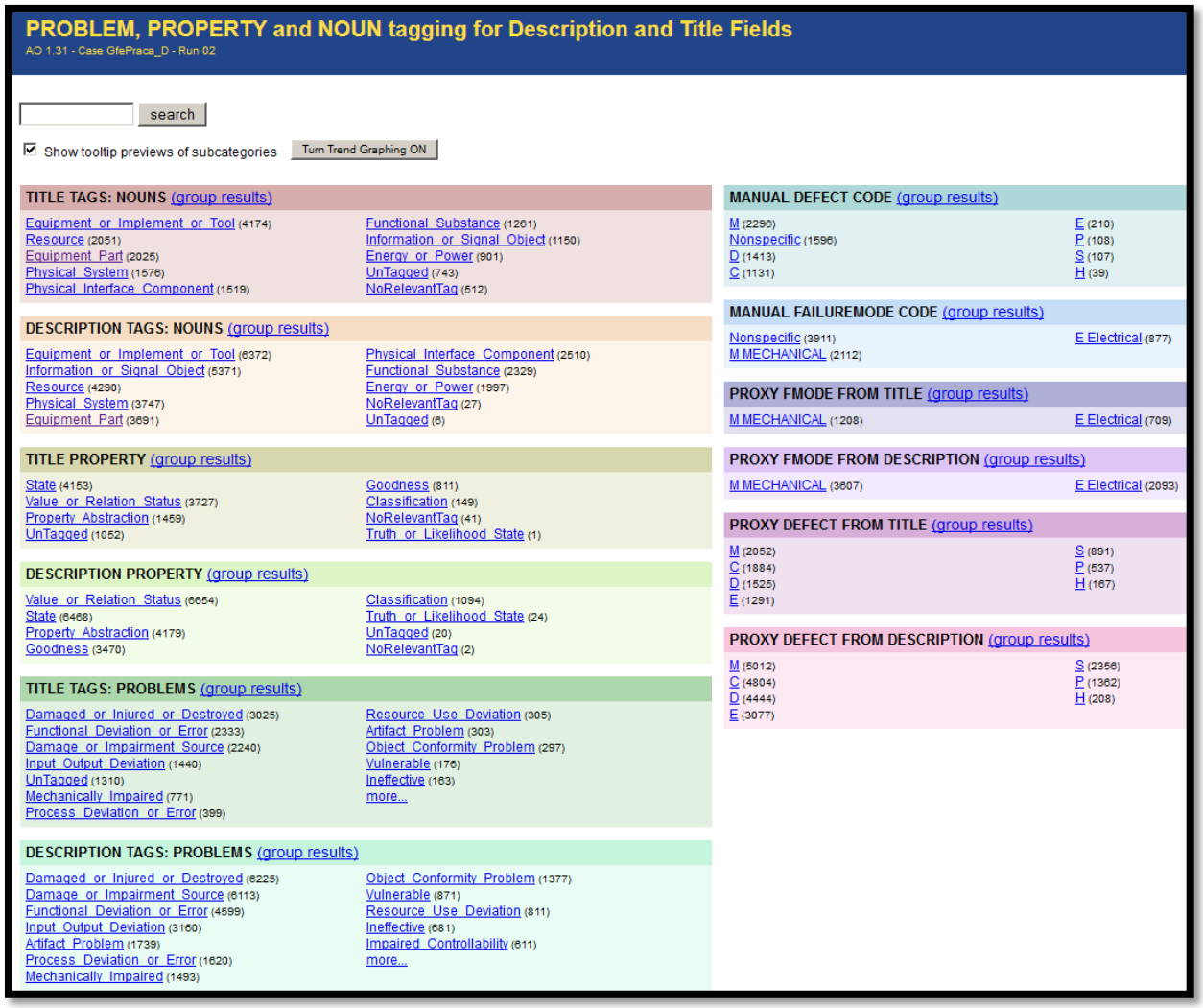


Figure E-19. Initial View of GFE PRACA Flamenco Data

The labels for the data categories have been abbreviated for the display. Below is an explanation of those abbreviations:

- Title tags: nouns
    - Title field of the merged record was processed to identify concept tags.
    - The noun portion of the Aerospace Ontology was used for concept tags for things like equipment.
- Description tags: nouns
    - Description field of merged record.

- Title property
    - Title field of merged record.
    - Property portion of Aerospace Ontology – states and characteristics.
- Title tags: problems
    - Title field of merged record.
    - Problems portion of Aerospace Ontology – anomalies.
- Manual Defect Code – original entry by reporter of the incident
- Proxy Fmode from Title
    - Proxy code (from rule defining construction of proxy code from concept tags).
    - Fmode – failure mode proxy code.
    - Title field of merged record was processed to derive proxy code.

For this scenario, the discipline expert is searching for information related to "inadvertent locking of TRRJ/SARJ DLA while the joint is rotating. The remainder of this illustrated scenario show some steps the can be taken with Flamenco to view this information. Initially, "joint" is entered into a keyword search (see Figure E-20).

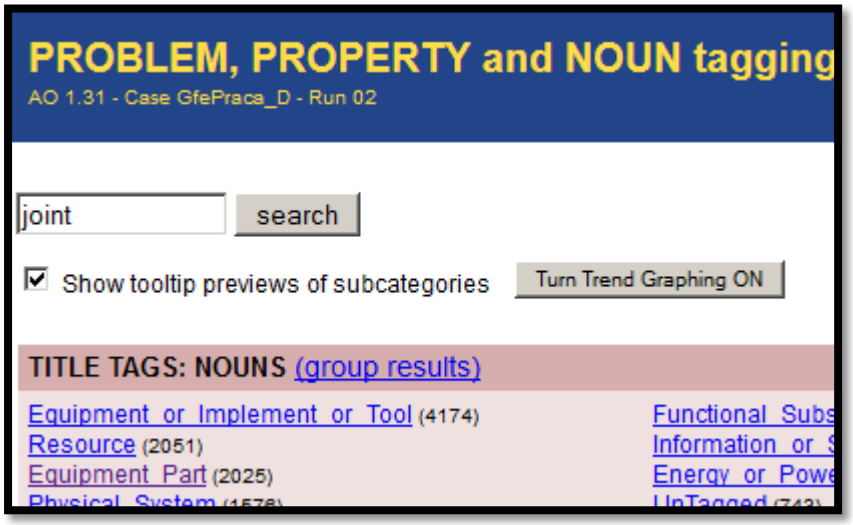


Figure E-20. Specifying a Keyword Search in Flamenco

Figure E-21 shows the search result.

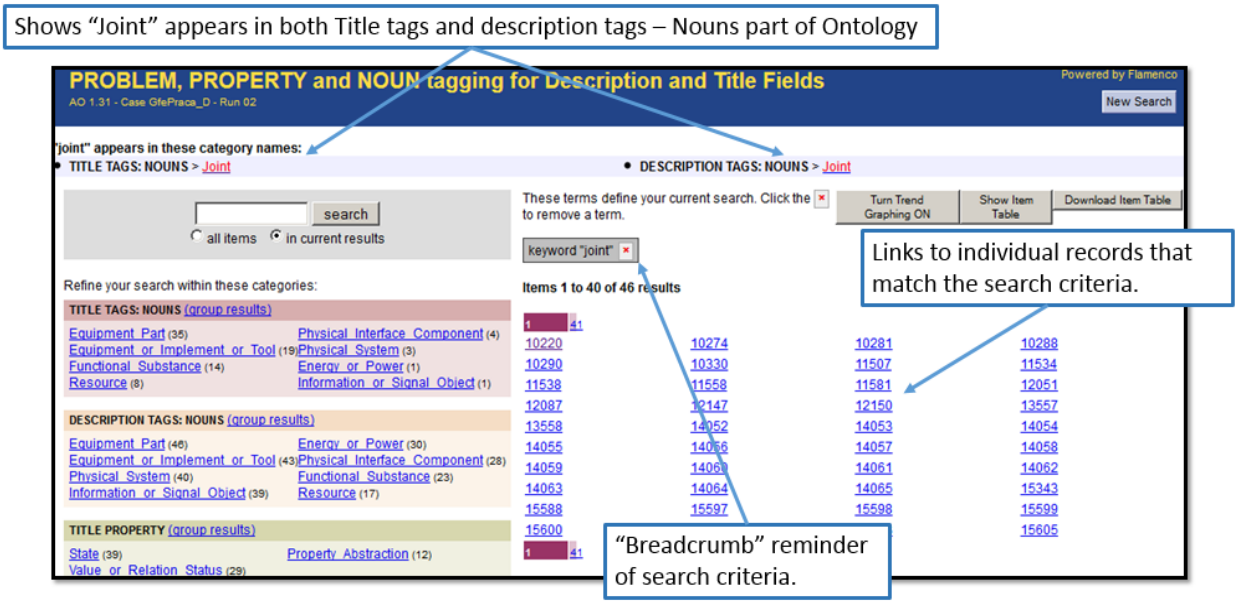


Figure E-21. Results of a Flamenco Keyword Search

A second keyword search is entered (see Figure E-22).

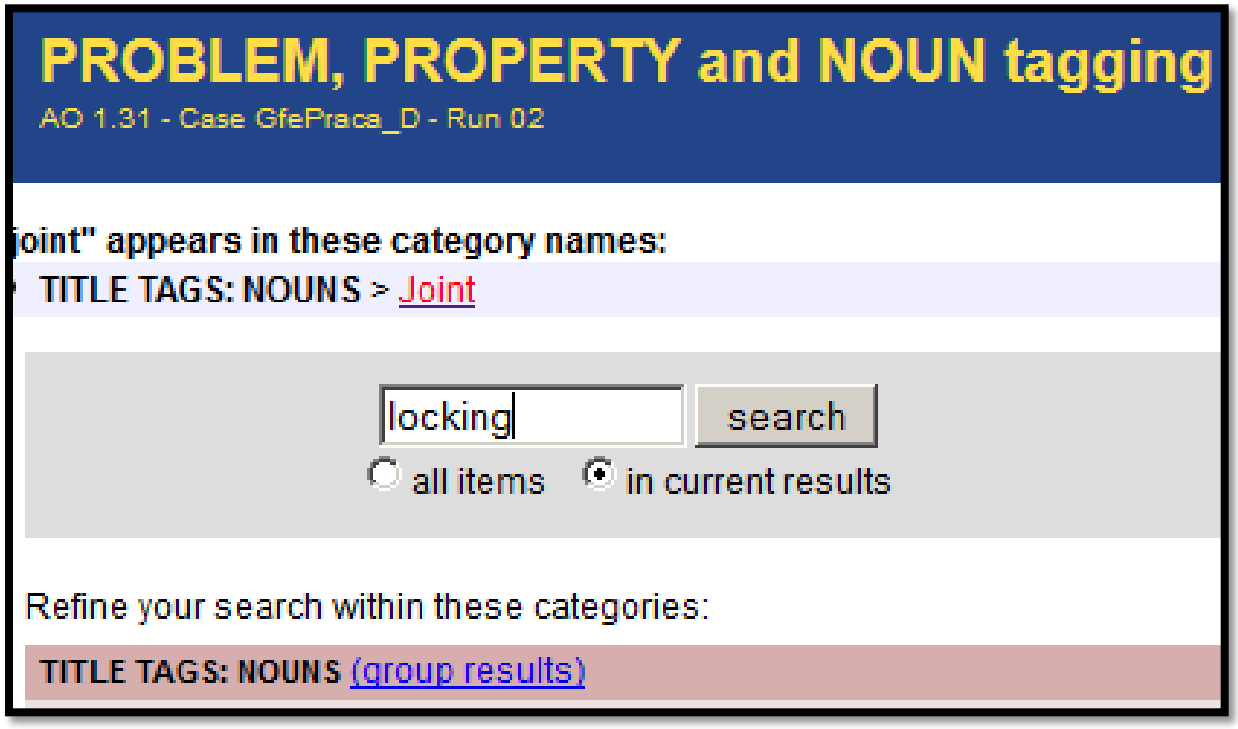


Figure E-22. Adding Another Keyword Search in Flamenco

Figure E-23 shows these results.

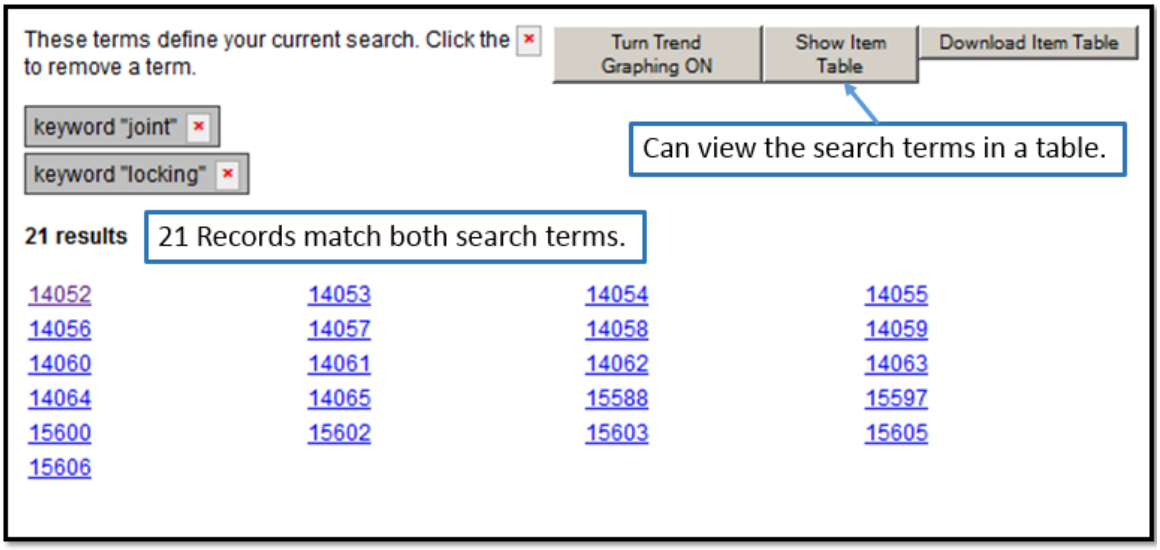
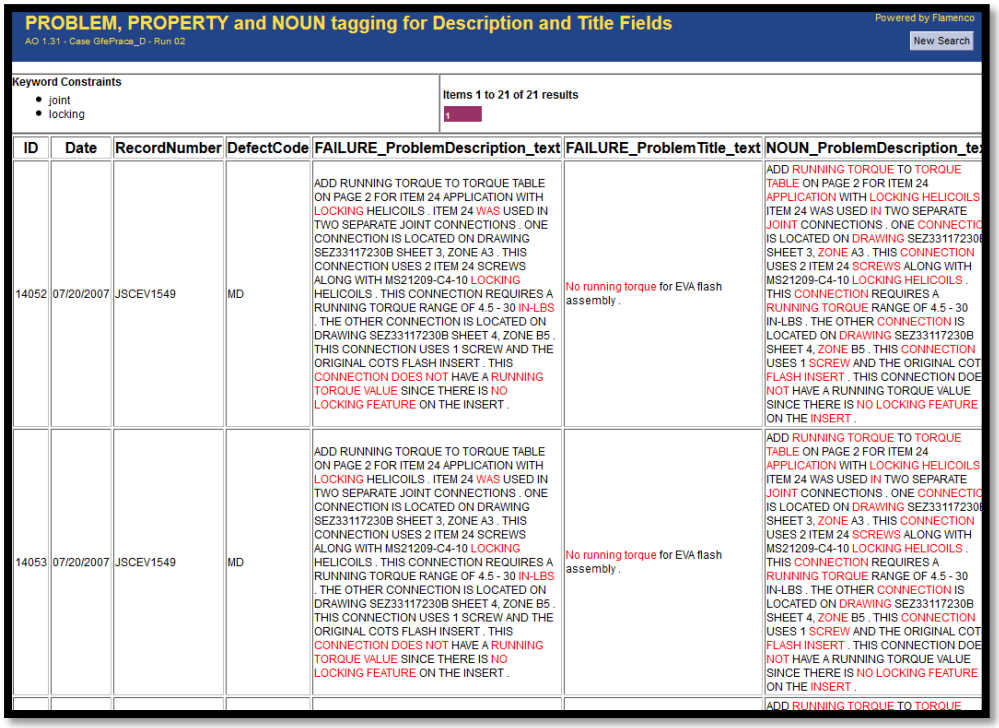


Figure E-23. Results of Two-keyword Searches

Figure E-24 shows a table view of the records matching "joint" and "locking."



ID	Date	RecordNumber	DefectCode	FAILURE_ProblemDescription_text	FAILURE_ProblemTitle_text	NOUN_ProblemDescription_tex
14052	07/20/2007	JSCEV1549	MD	ADD RUNNING TORQUE TO TORQUE TABLE ON PAGE 2 FOR ITEM 24 APPLICATION WITH LOCKING HELICOILS . ITEM 24 WAS USED IN TWO SEPARATE JOINT CONNECTIONS . ONE CONNECTION IS LOCATED ON DRAWING SEZ33117230B SHEET 3, ZONE A3 . THIS CONNECTION USES 2 ITEM 24 SCREWS ALONG WITH MS21209-C4-10 LOCKING HELICOILS . THIS CONNECTION REQUIRES A RUNNING TORQUE RANGE OF 4.5 - 30 IN-LBS . THE OTHER CONNECTION IS LOCATED ON DRAWING SEZ33117230B SHEET 4, ZONE B5 . THIS CONNECTION USES 1 SCREW AND THE ORIGINAL COTS FLASH INSERT . THIS CONNECTION DOES NOT HAVE A RUNNING TORQUE VALUE SINCE THERE IS NO LOCKING FEATURE ON THE INSERT .	No running torque for EVA flash assembly .	ADD RUNNING TORQUE TO TORQUE TABLE ON PAGE 2 FOR ITEM 24 APPLICATION WITH LOCKING HELICOILS ITEM 24 WAS USED IN TWO SEPARATE JOINT CONNECTIONS . ONE CONNECTIO IS LOCATED ON DRAWING SEZ331172308 SHEET 3, ZONE A3 . THIS CONNECTION USES 2 ITEM 24 SCREWS ALONG WITH MS21209-C4-10 LOCKING HELICOILS . THIS CONNECTION REQUIRES A RUNNING TORQUE RANGE OF 4.5 - 30 IN-LBS . THE OTHER CONNECTION IS LOCATED ON DRAWING SEZ33117230B SHEET 4, ZONE B5 . THIS CONNECTION USES 1 SCREW AND THE ORIGINAL COT FLASH INSERT . THIS CONNECTION DOE NOT HAVE A RUNNING TORQUE VALUE SINCE THERE IS NO LOCKING FEATURE ON THE INSERT .
14053	07/20/2007	JSCEV1549	MD	ADD RUNNING TORQUE TO TORQUE TABLE ON PAGE 2 FOR ITEM 24 APPLICATION WITH LOCKING HELICOILS . ITEM 24 WAS USED IN TWO SEPARATE JOINT CONNECTIONS . ONE CONNECTION IS LOCATED ON DRAWING SEZ33117230B SHEET 3, ZONE A3 . THIS CONNECTION USES 2 ITEM 24 SCREWS ALONG WITH MS21209-C4-10 LOCKING HELICOILS . THIS CONNECTION REQUIRES A RUNNING TORQUE RANGE OF 4.5 - 30 IN-LBS . THE OTHER CONNECTION IS LOCATED ON DRAWING SEZ33117230B SHEET 4, ZONE B5 . THIS CONNECTION USES 1 SCREW AND THE ORIGINAL COTS FLASH INSERT . THIS CONNECTION DOES NOT HAVE A RUNNING TORQUE VALUE SINCE THERE IS NO LOCKING FEATURE ON THE INSERT .	No running torque for EVA flash assembly .	ADD RUNNING TORQUE TO TORQUE TABLE ON PAGE 2 FOR ITEM 24 APPLICATION WITH LOCKING HELICOILS ITEM 24 WAS USED IN TWO SEPARATE JOINT CONNECTIONS . ONE CONNECTIO IS LOCATED ON DRAWING SEZ331172308 SHEET 3, ZONE A3 . THIS CONNECTION USES 2 ITEM 24 SCREWS ALONG WITH MS21209-C4-10 LOCKING HELICOILS . THIS CONNECTION REQUIRES A RUNNING TORQUE RANGE OF 4.5 - 30 IN-LBS . THE OTHER CONNECTION IS LOCATED ON DRAWING SEZ33117230B SHEET 4, ZONE B5 . THIS CONNECTION USES 1 SCREW AND THE ORIGINAL COT FLASH INSERT . THIS CONNECTION DOE NOT HAVE A RUNNING TORQUE VALUE SINCE THERE IS NO LOCKING FEATURE ON THE INSERT .
						ADD RUNNING TORQUE TO TORQUE

Figure E-24. Table View of Flamenco Search Results

Figure E-25 shows the full view of one of those records.

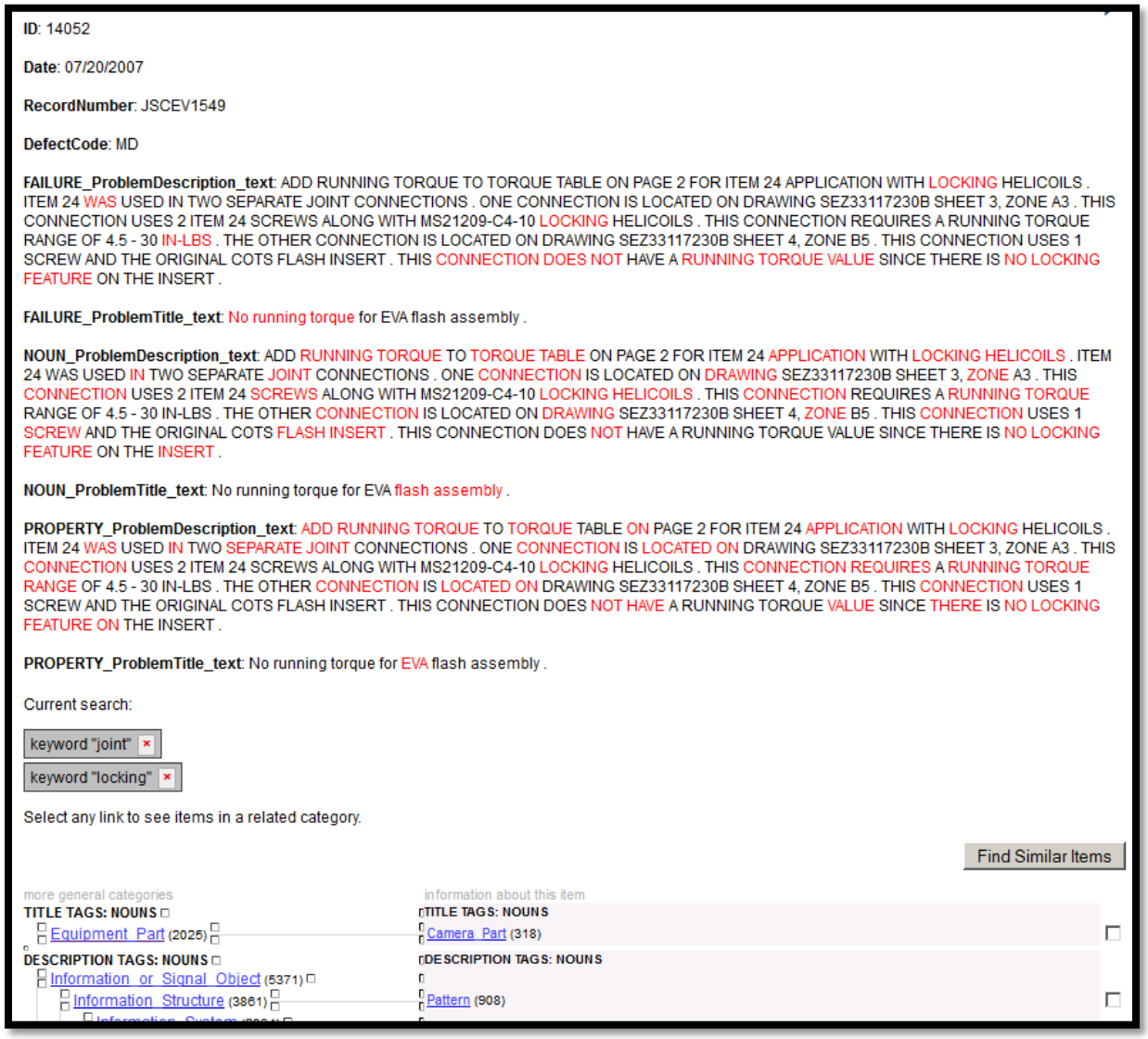


Figure E-25. Detailed View of a Flamenco Merged Data Record

The full view of the record contains the original text of the record with phrases that match concept tags highlighted in red. At the bottom of the display is a hierarchical view of the concepts related to those phrases (see Figure E-26).

Figure E-26. Hierarchical Concept Tags for a Flamenco Record

By checking the box next to "mechanically impaired," the analyst is able to look at those items related to that part of the Aerospace Ontology.

Figure E-27 shows the links to 14 records related to "joint" and "mechanically impaired."

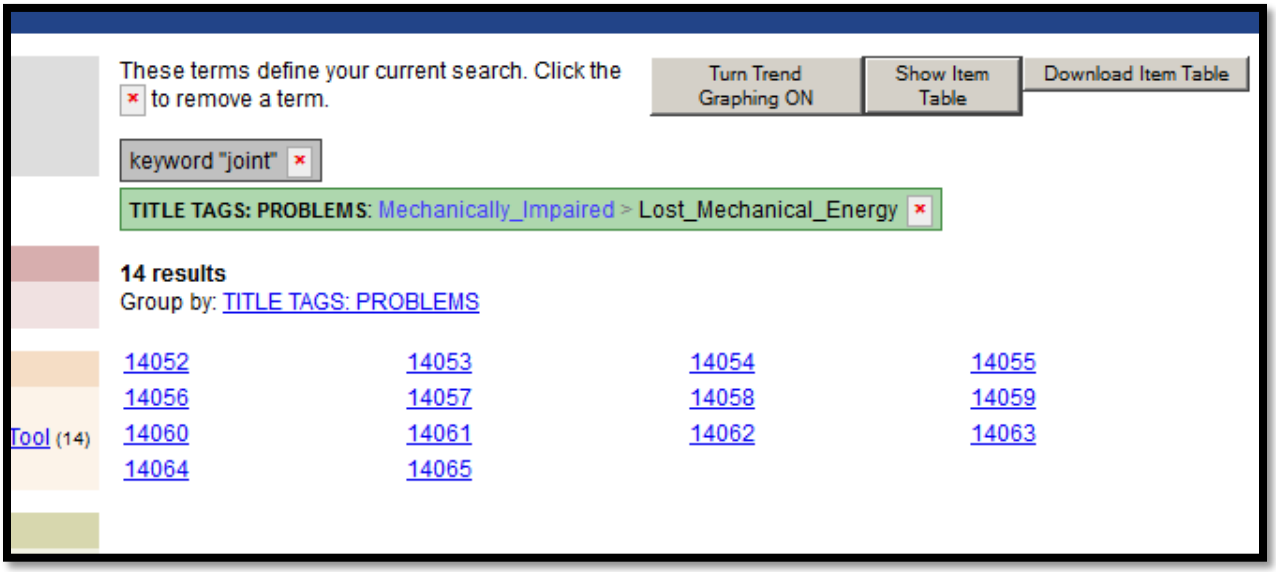


Figure E-27. Flamenco Search Results for Keyword and Concept Tag Combination

Finally, a table view of these records is displayed (see Figure E-28).

Facet Constraints

- **TITLE TAGS: PROBLEMS** = "Lost_Mechanical_Energy"

Keyword Constraints

- joint

Items 1 to 14 of 14 results

ID	Date	RecordNumber	DefectCode	FAILURE_ProblemDescription_text	FAILURE_ProblemTitle_text	NOUN_ProblemDescription_text
14052	07/20/2007	JSCEV1549	MD	ADD RUNNING TORQUE TO TORQUE TABLE ON PAGE 2 FOR ITEM 24 APPLICATION WITH LOCKING HELICOILS . ITEM 24 WAS USED IN TWO SEPARATE JOINT CONNECTIONS . ONE CONNECTION IS LOCATED ON DRAWING SEZ33117230B SHEET 3, ZONE A3 . THIS CONNECTION USES 2 ITEM 24 SCREWS ALONG WITH MS21209-C4-10 LOCKING HELICOILS . THIS CONNECTION REQUIRES A RUNNING TORQUE RANGE OF 4.5 - 30 IN-LBS . THE OTHER CONNECTION IS LOCATED ON DRAWING SEZ33117230B SHEET 4, ZONE B5 . THIS CONNECTION USES 1 SCREW AND THE ORIGINAL COTS FLASH INSERT . THIS CONNECTION DOES NOT HAVE A RUNNING TORQUE VALUE SINCE THERE IS NO LOCKING FEATURE ON THE INSERT .	No running torque for EVA flash assembly .	ADD RUNNING TORQUE TO TORQUE TABLE ON PAGE 2 FOR ITEM 24 APPLICATION WITH LOCKING HELICOILS . ITEM 24 WAS USED IN TWO SEPARATE JOINT CONNECTIONS . ONE CONNECTION IS LOCATED ON DRAWING SEZ33117230B SHEET 3, ZONE A3 . THIS CONNECTION USES 2 ITEM 24 SCREWS ALONG WITH MS21209-C4-10 LOCKING HELICOILS . THIS CONNECTION REQUIRES A RUNNING TORQUE RANGE OF 4.5 - 30 IN-LBS . THE OTHER CONNECTION IS LOCATED ON DRAWING SEZ33117230B SHEET 4, ZONE B5 . THIS CONNECTION USES 1 SCREW AND THE ORIGINAL COTS FLASH INSERT . THIS CONNECTION DOES NOT HAVE A RUNNING TORQUE VALUE SINCE THERE IS NO LOCKING FEATURE ON THE INSERT .
14053	07/20/2007	JSCEV1549	MD	ADD RUNNING TORQUE TO TORQUE TABLE ON PAGE 2 FOR ITEM 24 APPLICATION WITH LOCKING HELICOILS . ITEM 24 WAS USED IN TWO SEPARATE JOINT CONNECTIONS . ONE CONNECTION IS LOCATED ON DRAWING SEZ33117230B SHEET 3, ZONE A3 . THIS CONNECTION USES 2 ITEM 24 SCREWS ALONG WITH MS21209-C4-10 LOCKING HELICOILS . THIS CONNECTION REQUIRES A RUNNING TORQUE RANGE OF 4.5 - 30 IN-LBS . THE OTHER CONNECTION IS LOCATED ON DRAWING SEZ33117230B SHEET 4, ZONE B5 . THIS CONNECTION USES 1 SCREW AND THE ORIGINAL COTS FLASH INSERT . THIS CONNECTION DOES NOT HAVE A RUNNING TORQUE VALUE SINCE THERE IS NO LOCKING FEATURE ON THE INSERT .	No running torque for EVA flash assembly .	ADD RUNNING TORQUE TO TORQUE TABLE ON PAGE 2 FOR ITEM 24 APPLICATION WITH LOCKING HELICOILS . ITEM 24 WAS USED IN TWO SEPARATE JOINT CONNECTIONS . ONE CONNECTION IS LOCATED ON DRAWING SEZ33117230B SHEET 3, ZONE A3 . THIS CONNECTION USES 2 ITEM 24 SCREWS ALONG WITH MS21209-C4-10 LOCKING HELICOILS . THIS CONNECTION REQUIRES A RUNNING TORQUE RANGE OF 4.5 - 30 IN-LBS . THE OTHER CONNECTION IS LOCATED ON DRAWING SEZ33117230B SHEET 4, ZONE B5 . THIS CONNECTION USES 1 SCREW AND THE ORIGINAL COTS FLASH INSERT . THIS CONNECTION DOES NOT HAVE A RUNNING TORQUE VALUE SINCE THERE IS NO LOCKING FEATURE ON THE INSERT .
						ADD RUNNING TORQUE TO TORQUE

Figure E-28. Table View of Keyword and Concept Search Results

The analyst can continue to combine searches to find anomaly records of interest to find information related to the issue at hand.

Appendix F. Refining Proxy Codes

The proxy code rules were applied to the title and description fields of the GFE PRACA records. These rules were iteratively refined by comparing proxy codes to manual codes and their Help text definitions and then making adjustments to the rules. The rules were reapplied and reevaluated to see if the rule changes improved the correspondence between the proxy codes and the manual codes in the original GFE PRACA records. This process continued until diminishing returns were observed.

F.1 Early Iterative Refinements – Automatically Vetting and Refining During Proxy Code Assignment

Early iterative refinements were based on information retrieval statistics: comparing rules and their performance against manually assigned codes from the GFE PRACA source.

- Recall was defined as the percent of the total set of manual code examples where STAT assigned the same code. Proxy code recall performance was measured for those manual codes that have proxy code rules (i.e., eliminating noncommittal codes, obsolete codes, and codes with less than seven manual examples).
- Precision was approximated by the average number of proxy codes assigned to each record with a specific manual code.

F.2 Subsequent Proxy Code Refinement to Improve Usefulness to Discipline Experts – Vetting

The purpose of this round of refinement was to make the proxy codes more useful to discipline experts. The nature of the improvement depends on how the proxy codes will be used.

Considerations:

- Manual codes might be used in the first round of investigation by discipline experts. If so, the proxy code can help to address the question "What additional items should I see to investigate the current issue?"
- Recall asks, "What proportion of all records that should receive a given code ($t_p + f_n$, where f_n are false negatives) have been found (t_p – true positives)?" Formula: $t_p/(t_p + f_n)$
    - Manual codes may have a surprising amount of deficiency in recall because the coding scheme limits the coder to a single code for defect and for failure mode. For example, if each anomaly could reasonably be assigned three failure mode codes, 33 percent would be the highest possible recall for each manual failure mode code. Descriptions of anomalies frequently refer to several problems occurring in the incident.

  - Preliminary estimates of proxy code recall were about 30 percent, similar to manual recall. However, the recall failures of the proxy codes are likely to be different from those of manual codes, so that a combination of proxy and manual codes should have better recall than just proxy or just manual.
  - Thus, it is expected that proxy codes can reasonably provide the support our experts will want from them: What additional records should I see beyond those with a given manual code?"

- Precision asks, "What percent of records with a given code assignment ($t_p + f_p$, where f_p are false positives) have been coded properly (t_p)?" Formula: $t_p/(t_p + f_p)$
  - Hopefully, manual codes have a high precision. Experts report the problems, so most of the time when they assign a code, it should be a correct code. Even so, there were some striking counterexamples, described in F.5.1.
  - The approximation of proxy code precision that was used was an underestimation. Given the likelihood of multiple possible manual codes, not all records that could reasonably be assigned a code were given that manual code.
  - So, if the manual codes have OK recall and much better precision, it might make sense to address the issue of "what have we missed" with the proxy codes. Discipline experts trying to find well-hidden records related to an anomaly type can afford to filter out a few false positives.
  - On the other hand, we do not want to overload an expert with many false positive proxy codes. This can make it onerous to wade through false positives for a few good example proxy codes. This is the motivation for vetting, or removing as many false positives as possible, thus improving precision.

To make the proxy codes more useful, we should consider taking measures to remove lower-precision proxy codes so that the expert has less noise to sift through. Removing false positive proxy codes is the purpose of vetting. Vetting options include:

1. Code by code, measure proxy precision, and **remove those proxy codes that have poor precision**—in other words, not show proxy codes that have precision below some threshold.
   - This would involve **manual examination of individual records**.
   - A sample from each proxy code value would be examined manually so that a precision can be computed to decide which proxies to suppress.
   - This examination would be for 26 sets of GFE PRACA description fields for failure mode codes and 31 sets of GFE PRACA title fields for defect codes. If we examine 20 records manually in each set, up to 1,140 records {(26 + 31) * 20 = 1,140} would need to be examined manually to establish a precision for each code.

- For each set of records, each manual code would need to pass a test of correctness so that any examples where a manual code was inappropriately assigned to a record would be eliminated.
- This option would allow assessment of manual coding precision with almost no additional work. During this development, it became clear that the manual codes should have been vetted. Given the low accuracy of some manual codes, the statistical machine learning approach (which was used to define rules for proxy defect codes) should not be used unless vetting of manual codes results in selection of accurate training sets.
- Strengths
  - Can eliminate many false-positive proxy codes, to improve usefulness to discipline experts.
  - Allows assessment of proxy coding precision. For example, for each proxy code, how many matched the vetted code?
  - Allows assessment of proxy recall. For example, for each vetted code, how many were matched by proxy codes?
  - Allows assessment of manual coding precision. For example, of the records examined in each manual category, how many manual codes were inappropriate?
  - Allows assessment of manual recall. For example, for each vetted code, how many were matched by manual codes?
  - Allows comparison of proxy and manual coding performance.
- Weaknesses
  - Sampling means there is no guarantee that no record contains large numbers of false positive proxy codes.
  - Some number of data base codes will have no proxies assigned.
  - Manually examining 1,140 records and assigning all the codes that apply is laborious and costly.
  - Multiple raters will be needed to ensure that the gold standard set of manually coded records is good enough.

2. For each record, **suppress all proxy codes for any field with too many proxy codes** (e.g., if a field has more than five proxy codes, do not show any proxy codes for that field).
   a. This option **does not require an exhaustive manual examination** of 1,140 records to make a good precision measure.

b. The number of assigned codes is not a pure precision measure, but it still addresses the general precision concept.

c. Strengths

i. Guarantees that no record contains a large number of proxy codes.

ii. Less likely that discipline experts will be flooded with false positives for a given proxy codes search.

iii. Can be implemented with software—no manual assessment required—less laborious and cheaper.

d. Weaknesses

i. No guarantee that a discipline expert search will never result in lots of false positives.

ii. Does not develop a gold standard to assess precision and recall of proxy and manual codes.

3. A **variation** on the first approach would be to eliminate proxy codes that do not have good precision as measured by matching manual codes. The NESC assessment team took initial steps to arrange the data to answer the question, "For each proxy code, how many matched manual codes?"

4. A **variation** on the second approach would be to manually examine the fields in records with too many proxy codes, to eliminate the incorrect proxy codes, leaving a smaller subset. This could involve examining many records, but the quantity is likely to be less than 1,140 records.

F.3 Decision: Proxy Code Precision Improvement Measure

The NESC assessment team decided to identify those records with more than five proxy codes and retain only those proxy codes that scored the highest precision in the initial measure of proxy code precision (agreement with manual codes). This prevents the overloading of discipline experts with records of more than five proxy codes (so there should be fewer false positives to sift through).

F.4 Proxy Code Performance Assessment

The limited human resources affecting the refinement of proxy codes and their rules also influence decisions about the assessment of proxy code performance. Ideally, a team of four to five judges would sift through the records and identify a set of 20 records for each proxy code that can be agreed as exemplars of each possible trend code. From the computations above, that would mean looking at over 1,100 records with a team of four or five members. With such an assessment in hand, it would be straightforward to measure both recall and performance of both manual and proxy codes.

It is important to keep in mind that the **recall** is the more important performance measure because the use case for proxy codes is expected to address the question, "What additional records should I examine to ensure that I have not overlooked important cases?"

Precision can be less exhaustively measured. The manual vetting team can look at 20 exemplars of each code value assigned and see how many appear to be appropriate. That is a pretty good estimate of precision (of the codes assigned a given value, how many were appropriate?). It would be good to compare performance of proxy and manual codes.

Recall is much more difficult to do in a cost effective manner. The recall question is, "Of the records that should have received a given code, how many were assigned that code?" Ideally, a subset of records would be examined by a team of manual vetters to find 20 clear examples of each database code. This set of 20 would be the denominator for the recall (records that should have received a given code).

- Option 1: Look through records until less than 20 clear exemplars are found for each database code. We might aim for 15 for the more frequently used codes and 7 for less frequently used codes.
    - Still a large set.
    - May be expensive.
    - Values for less frequently used codes would still be unstable.
    - Much better than using manual codes as the standard.
    - Both recall and precision can be computed for both manual and proxy codes from the same set of records.
- Option 2: Look through the records until a smaller set of clear code exemplars are found, regardless of how many representatives of each code value have been identified. In this case, a smaller total set can be examined because the goal is only to derive an overall recall score, not a score for each code value.
    - Find the first set of perhaps 200 code exemplars, regardless of the number of exemplars for each code value. Only keep clear exemplars for this set and only consider the specific codes that are not noncommittal or obsolete codes.
    - The number of exemplars may need to be reduced to accommodate what the project can afford.
    - Use that set to measure recall, for both manual and proxy codes, so that there is some notion of the performance of both sets.
    - This is far less manual vetting work.
    - It provides a good overall estimate of coding recall for both manual and proxy codes.

  - Precision would be computed by assessing how many of the assigned codes were appropriate. Thus, recall and precision would be computed on a different set of records.

- Option 3: Look through the manual codes (eliminating noncommittal and obsolete codes) to find a set of 200 clear exemplars (accurate manual codes). Then, look to see how many were matched by proxy codes.
  - Again, the number of exemplars may need to be reduced to accommodate what the project can afford.
  - This is probably less manual work than option 2.
  - It does not allow comparing manual recall to proxy code recall performance. This leaves no good comparison of recall performance for proxies.
  - It does allow a single estimate for proxy code performance.
  - Precision would be computed by assessing how many of the assigned codes were appropriate. Thus, recall and precision would be computed on a different set of records.

F.5 Inadequacy of Manual Condition Codes

The effectiveness of types of searching, browsing, indexing, classifying, and coding depends on the type of analysis they are used for. Does the scheme of codes and retrieval strategies help analyze recurrences and trends? Does it increase recall, by finding all instances of a specific type of problem, like a crack? Do the fields and codes represent distinctions with a difference? For example, does an action like a corrective response depend on a field and code distinction, like type of failure mode or type of defect? Does it help exploration for unanticipated types of problems? Does it help find root causes or contributing factors?

The purpose of manual condition codes is to extract reduced information from reports, to locate or select more relevant reports and gather groups of reports with common conditions needed for analysis (i.e., failure modes, defects, causes). These codes provide one way of overcoming some weaknesses of full-text search: synonyms; variants such as abbreviations, acronyms, and misspellings; and homonyms (i.e., terms with multiple meanings). The manual coder can easily interpret all these variations while identifying a code from a standard set that best fits the report. Then analysts can focus on specific fields and codes to guide retrieval of a relevant item or group of items.

Manually assigned condition codes were included in some of the databases in the data set. They were included in fields with codes for types of failure modes, defects, and causes. The coding schemes made merging data difficult because they were not standard across the data sets. Coders would sometimes not assign a code to a field, and there were opportunities to assign nonspecific codes. Establishing identical trend code fields across data sets could help standardize

information retrieval. Permitting more than one code per report would overcome the limitation of allowing only one code to be assigned to each field. This also enabled supplementing nonspecific codes with relevant specific codes.

Rules were developed to assign "proxy" codes to condition fields, either by manual inspection or statistical machine learning. The rules used the ontology-based concepts that were extracted from title or description fields in each data record, by semantic text analysis. The rules tested logical combinations of the presence or absence of a concept tags associated with a record. Some examples of these rules, using OR logic applied to concepts associated with failures, are included in Table 6.4.4.1-1.

F.5.1 Observed Problems

It was assumed that the manual code assignments were a good basis for developing the rules and evaluating the accuracy of the proxy codes. However, this was not a safe assumption. Serious manual coding errors were found during the process of developing the proxy code rules. In the GFE PRACA and PART PRACA data sets, manual coding errors in the fields for failure mode codes and defect codes were much worse than expected. Manual coders who misapplied codes appeared to either not understand or not read the Help text for these codes. Both cases were observed. For example:

- At least 173 of the 195 GFE PRACA records that have manual MD failure mode codes (delayed or slow operation) appear to be manually miscoded. All 173 concern peeling heat shrink, which seems unrelated to delayed or slow operation.
- Code confusion errors were common among these code pairs: Fails Off vs. Fails On; Fails Closed or Fails to Open (Extend) Completely vs. Fails Open or Fails to Close (Retract) Completely. Eleven examples of these errors in PART PRACA code assignments are shown in Table F-1.

There can be high variability in agreement on specific codes. A few or many codes can be misinterpreted or misused. There are also multiple possible types of coding errors that result in incorrect assignments and low precision:

- Misinterprets code definition or unable to fill in gaps in short definition.
- Misinterprets how to assign codes to multiple condition fields, especially when some overlap.
- Misinterprets description/report or unable to fill in gaps in report.
- Chooses nonspecific code.
    - Varying reluctance to commit to specific code.
    - Appropriate code not found in set.
- Uses only a subset of codes.
- Copies a code from a related report (which may be incorrect). This may be the cause of the MD code errors that were observed.

Table F-1. Example of Manual Code Confusion from PART PRACA

Comment: GFE PRACA Help Text	Proxy Code Rule	# Records with code	Suspect Coding Examples	Suspect Coding Examples	Suspect Coding Examples	Suspect Coding Examples
Fails Off	Fail_Off | Electrically_Disconnected | Electrical_Imbalance	61	FULL LEVEL Z-AXIS RANDOM VIBRATION TESTS WEREPERFORMED WITH THE TDLA LAUNCH LOCK PIN NOT ENGAGED .			
Fails On	Fail_On	26	During pre-Z Axis Vibration to -6 db per E039447 Rev 3 and E040377 Rev 9, at approximately 1.5 minutes into -6 db attenuation . IS : FI tripped openSB : Closed with UUT operating in Standby Mode 2 .	While inbetween setups for RS03 testing the OIV / NIV valve waspowered down . When the valve was powered up again the LED `sfailed to light up and the valve failed to actuate whencommanded .	Unit did not turn on after Y-AxisVibration Testing of Para 8.12.0 of Qualification TestProcedure MP6495 .	
FAILS CLOSED OR FAILS TO OPEN (OR EXTEND) COMPLETELY	Fail_Closed | Fails_to_Open | Did_Not_Deploy	16	Activated PCUs 1 & 2 for Ionosphere data collection . At termination of data collection period Plasma Contactor Unit-2 (PCU-2) failed to transition to Shutdown Mode . IFI MER-01458 , 05 APR 04, was opened and it was determined that PCU-2 Latch Valve # 05 (LV1)- Failed Open .	DURING THE POST " Y " VIBRATION FUNCTIONAL TEST THE SQUARE GRID INTERFACE (SGI) CLAMPING MECHANISM DID NOT OPERATE PROPERLY . THE LEFT SIDE DID NOT RETRACT WHILE RUNNING STEP 4.4.3.11 OF THE ...TEST PROCEDURE .	After switch is activated , switch will not go off whenreturned to steady state .	
FAILS OPEN OR FAILS TO CLOSE (OR RETRACT) COMPLETELY	Fail_Open | Fails_to_Close | Did_Not_Retract | Fails_to_Turn_On | Underreactive_Function | Stuck	28	During thefirst 20 F dwell of para 4.2.10g qualification test thevalve failed to open when commanded to do so .	Visual indications during the rendezvous show that the SM Port Solar Array 2 is not fully deployed . One panel appears to be folded back at 90 degrees .	Failed detail functional test segments9, 10, 13- 15, 18, 19, 22 during 24th cycle (last cycle) ofthermal cycling at low temperature (10 deg C) . The failuremode was loss of Audio Channel 0 output .	During the acceptance thermal cycle test , one of thePressure Control Panels (PCP) , Oxygen / Nitrogen IsolationValves (OIV / NIV) failed to actuate from the close to openposition when commanded during the functional test . ...

F.5.2 Quality Criteria and Error Sources

Criteria for the quality of coding schemes include utility/applicability, clarity, reproducibility, and difficulty.

Utility and applicability concern the relevance of the coding scheme to analyses. Previously defined codes and fields may not support analysis of events and concerns that come up in a program. For example, the current manual condition coding schemes are not likely to make it easy to analyze the following specific cases:

- Spontaneous resets in processors in a power system, causing power cycling in powered equipment.
- Any cracks that have happened in vehicle and its systems.
- Failure modes that are associated with aging and end of life.

Clarity is achieved with a well-defined and distinct fields and codes. Criteria for belonging to a class/code or group need to be well defined and complete. They can be ambiguous if they are abbreviated or the criteria can be expressed in language that is not aligned with the language of the reports. If they are too short, they can be unclear because of missing examples or detail. If the coder is constrained to select a single code and no secondary codes are allowed, then guidance is needed as to what characteristics should be primary or preferred in assigning the

code. The anomaly condition codes are generally not well defined because the Help text is brief and often confusing, as shown in Table F-1. No guidance is given on what code assignment should be used when multiple alternative codes are possible.

For clarity, fields and codes need to be distinct and consistent. The anomaly condition codes have some of the following weaknesses:

- Overlaps between classes/codes or fields without guidance on how to handle.
- Large and complex multilayered code sets.
- Inconsistent structure of fields and codes.
    - Types of relations between the concepts/fields are not explicit or well defined.
    - Subtype-supertype relations are mixed with other relations in code hierarchies, violating the assumption that all the characteristics of the superset are applicable for the members of the subset.

Difficulty is affected by data overload or inadequate data. In a large set of possible code assignments (fields × codes), it is easier to overlook a relevant coding rule or miss a key characteristic of an anomaly. On the other hand, missing coding information in the report can lead to assigning a nonspecific code or assigning what would have been a secondary code.

Reproducibility is frequently measured as inter-rater reliability between two or more coders. Coders agree on the code assignment, which is more than agreeing that a field in a report could be assigned that code. While it may be easy to rule out many possible code assignments, there may only be fair positive agreement on the assignment selected from the remaining codes. Percent agreement is the simplest and most intuitive metric. Other metrics take into account the amount of agreement that could be expected to occur by chance.

Common causes for low reproducibility, beyond problems with clarity and difficulty, include:

- New context and its associated issues may require some shoehorning of partially matching codes.
- Personal and local interpretations, coding guidelines or procedures.
    - Facility-specific or discipline-specific priorities that differ from the guidelines.

F.5.3 Remediation and Recommendations

The primary remedy for coding errors includes effective procedures for development, review, and update of coding schemes. A second line of defense is training and help, such as advice and additional information in FAQs. A third area of remediation is application of the time and resources needed to do the reporting and coding tasks, so that quality does not suffer from shortcuts. All of these can require significant investments.

Measuring manual code error rates would help identify the subset that could be used for auto-generation of proxy code assignment rules by machine learning (which was used to define rules for Defect Code proxies). Without this information, the proxies for the defect codes are of

unknown quality. Likewise, a subset of failure mode codes with good inter-rater reliability on a subset of reports could be identified and used for manual development of proxy code assignment rules for the failure mode field.

Are these strategies enough to remedy the problems with manual code in the merged data set? Remediation strategies do not overcome utility problems. In this study, manual condition codes were not found to be productive for the analyses because they did not help much in identifying groups of reports that are relevant to new issues that came up in the program. Other assumptions need to be revisited. GFE and MOD AR concerns are very different. Would interoperability really be achieved by applying proxy rules based on GFE PRACA codes to the other databases?

Appendix G. SAS® Analysis with Text-Mining Topics

The purpose of the SAS® analysis text-mining phase was to find reports in certain problem areas, disciplines, or subsystems that could not be found easily with keyword search. Technical discipline experts specified lists of terms and noun groups that defined areas of focus. Statistical text mining was used to identify correlated documents, based on terms and noun groups they had in common. Each group of correlated documents represents a latent topic, which is defined by the common terms. Thus, new terms or noun groups could be identified to add to search expressions. The analysis was used to determine significant observations or trends that needed further investigation.

During the analysis phase, reducing the noise then became the focus. "Noisy" terms do not contribute to correlational analysis and thus do not help to discriminate between documents in text mining. To reduce noise in the analysis, the Text Parsing and Text Filter node properties were modified. The Text Parsing Node was changed to ignore most standard types of Entities, because they are not used in text fields in anomaly reports. The blue Entity types in Figure G-1 were ignored. More Stop Lists terms were added as needed (as illustrated in Figure G-2).

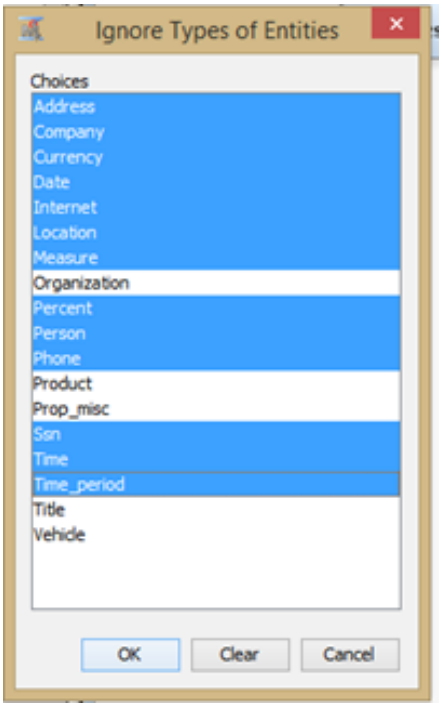


Figure G-1. Ignore Types of Entities

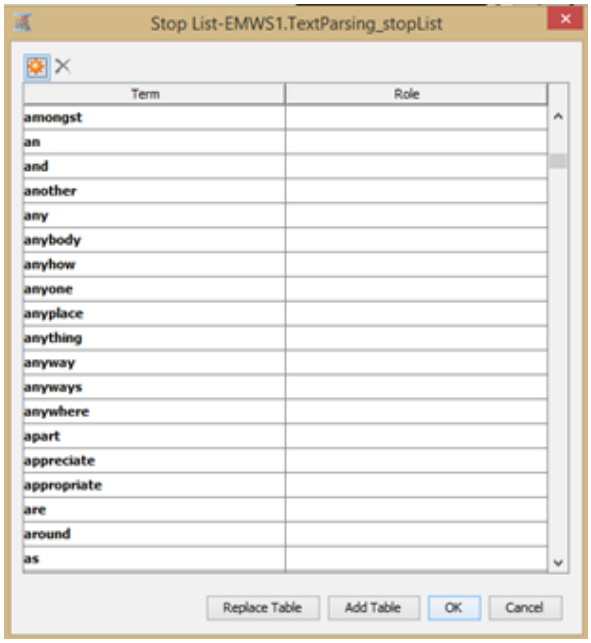


Figure G-2. Stop List Example

For efficient analysis, the Text Filter node is most important in this phase. The main purpose of the Text Filter Node is to weight terms based on their importance in the corpus of data records. Frequent terms are "noisy" in text mining, because they are not helpful in discriminating information in the documents. They receive very low weights. Common types of weighting settings in the Text Filter node property sheet are:

- Frequency Weighting (Local Weight) accounts for how terms relate within a document. Frequency weights such as Log and Binary are available.
- Term Weighting (Global Weight) accounts for how a term is spread across the corpus. A number of term weighting methods are available such as Entropy, Inverse Document Frequency, and Mutual Information.

These weights are used for more effective dimension reduction in later processing by the SAS® Text Miner. Dimension reduction is a way of reducing noise while keeping enough information to represent the original data. Text Miner uses Singular Value Decomposition (SVD) matrix factorization for dimension reduction.

A balance is necessary when looking for trends. Some trends might not be seen because the noise reduction setting is too high. The approach for this assessment was to initially allow for a higher level of noise. After review of the data, a change in weights was made to focus on more specific areas. A given corpus can react directly to these node options. It was important to experiment with weighting methods to find optimal settings.

The Text Filter node is also used for data cleaning, term exploration, and querying (see Figure G-3). The Minimum Number of Documents was changed from 1 to 4. Spell Check and Filter Viewer properties were also used. It was possible to create a Search Expression to filter documents, to focus on target areas that discipline experts requested or other areas of interest that were observed during the analysis (see Figure G-3).

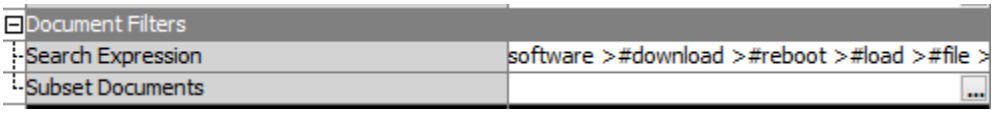


Figure G-3. Text Filter Node Search Expression

The Text Filter node for each discipline term (see Table G-1) was added to the search expression (see Figure G-3) to focus on requests from software, human factors, electrical discipline experts, and others. Each discipline requested specific areas of focus terms or terms and noun groups that were discovered during the analysis and were added to the search expression and rerun.

Table G-1. Terms Specified in Discipline Requests

Text Mining Analysis Request	
Disciplines/Subsystems	**Terms and Noun Groups of Interest**
Software	software
	download
	reboot
	load
	file
	laptop
	ssc
	pc
	server
	network
	boot
	computer
	crash
	firmware
	command
	opslan
	windows
	oca
	service pack
	downlink
	screen
	cycle
	telemetry

Text Mining Analysis Request	
Disciplines/Subsystems	**Terms and Noun Groups of Interest**
	dump
	t61p
	signature
	uplink
	shell
	client
	transfer
	code
	encoder
	no joy
	backup
Human Factors	payload
	crew
Electrical	Elect
	electrical
	power
	current
	switch
	light
	alarm
	fault
	isolate
	wire
	trip
	circuit
	reset
	resistance
	batteries
	jumper
	power cycle
	spike
Thermal	thermal
	thermally
	heat
	tps
Special Request (Sensor and Transducer)	sensor

Text Mining Analysis Request	
Disciplines/Subsystems	**Terms and Noun Groups of Interest**
	transducer
	thermistor
	valve
	lvt
	actuator
Special Request (pump)	pump

Using the Text Filter Node with the Search Expression properties, SAS® selected a subset of documents. Each of the nodes and properties were adjusted on a trial and error basis to get to the right level of information necessary. After each run of data through the SAS® Text Miner process, properties were adjusted to provide meaningful and understandable information for the discipline experts.

The user interface provides a Text Filter Snippet (see Figure G-4) for examining where the terms are being used within each document.

TEXTFILTER_SNIPPET
... was downlinking CHECS **files** at the time . At ...
... to satisfy the **load** request of 7250W . At ...
... EEPROM attempted to **load** S1-1 for PPL with new ...
... the Automated IWIS **software** . ...
... weekend , the **loads** would keep changing spontaneously . ...
... the 1.4 GB **file** . The PI for the ...
... ODF . ZIP **files** . Log **files** are not ...
... command had been **loaded** but not Confirmed . ...
... capability to transfer **files** from the edls-edp **server** to ... to
... the ghosting of **PCS** hard drive S / N ... of the drive **loaded**
... charger is in **PCS** mode for maintenance cycles , ... up on the
... MSS checkpoint data **files** that were delivered to MCC ...
... observed the Airlock **PCS** disconnected and the **PCS** count ...
... Tool on the **SSC** client . ...
... (EV3) **files** 1 -50 . Due to ... the remaining EV3 **files** and
... 2 . DMC **Loaded** PEHG-2 Configuration 3 and cleared ...
... that the LAB **PCS** had failed . He power ... swap the CUP **PCS**
... Cygnus Overview crew **PCS** display . The units listed ...) and
... the . WMV **file** embedded in the MSWord document ... insalled
...) that the **SSC File Server** had locked up ... G . No **SSC** clients
... were no workstation **loading** issues due to command inventory
... The Crew has **rebooted** NGSD 8 multiple times . ...

Figure G-4. Text Filter Snippet Sample

The Text Filter Viewer also supports interaction with the data. Ignored terms are still part of the data set but have weights of 0.0. In the Filter Viewer, those ignored terms can be restored by adjusting the weight (see Figure G-5). Roles (e.g., parts of speech: nouns, verbs, etc.) addressed

the problem of terms with multiple meanings. This was vital for unstructured text fields when problem report initiators often used both roles as a verb and a noun for the same term. Other data sets sometimes benefitted from turning off "parts of speech" and using a "bag of words" approach to text mining that ignores word order and syntax. Both methods were used during text mining.

Terms

TERM	FREQ ▼	# DOCS	KEEP	WEIGHT	ROLE
be	26719	4517	☐	0.0	Verb
not	5701	2798	☐	0.0	Adv
gmt	4312	2103	☑	0.141	Miscellaneous Pro…
crew	4119	1608	☑	0.164	Noun
data	3344	1388	☑	0.193	Noun
s	3106	1517	☐	0.0	Noun
report	2895	1577	☑	0.162	Verb
n	2667	1113	☐	0.0	Noun
have	2627	1635	☐	0.0	Verb
file	2549	1064	☑	0.222	Noun
no	2270	1415	☐	0.0	Adv
do	2243	1399	☐	0.0	Verb
time	2077	1288	☑	0.18	Noun
that	2002	1382	☐	0.0	Adv
command	1992	708	☑	0.273	Noun
error	1989	940	☑	0.228	Noun
perform	1892	1262	☑	0.18	Verb
c	1769	648	☐	0.0	Noun

Figure G-5. Term Weights and Terms Ignored

The Text Filter Viewer supports identifying Synonyms to further reduce the noise in the data (see Figure G-6). For example, "computer" could be set as a synonym of "PC." Synonyms had to be assigned with caution. For example, "computer," "PC," laptop," and "desktop" could all be synonyms. However, this would obscure the data if a particular user was looking for problems that affected just laptops.

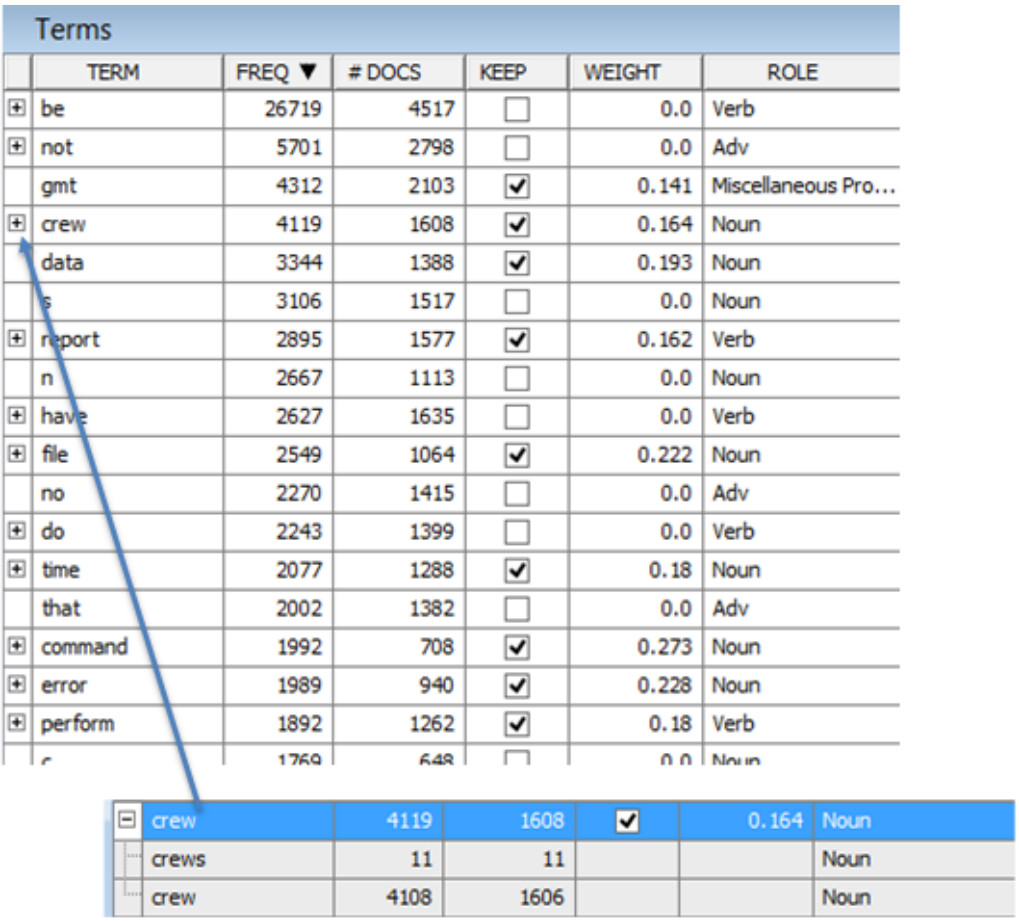


Figure G-6. Synonyms

The Text Topic node is the final node in this part of the analysis. Properties for this node determine how many topics are formed via correlational matrix analysis (see Figure G-7). This node enabled the exploration of problem report document collections by automatically associating terms and documents for both discovered ("latent") and user-defined topics. Topics are collections of terms that describe and characterize a main theme or idea in a set of related documents. The Text Topic node assigns scores that measure the association between each topic and each document and between each term and each topic. Thresholds determine whether the association is strong enough to assign the document or term to the topic (see Figure G-8). Documents and terms may belong to more than one topic or to none at all (see Figure G-9).

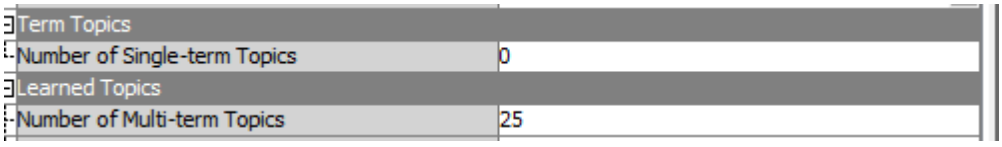


Figure G-7. Term Topic Properties Settings

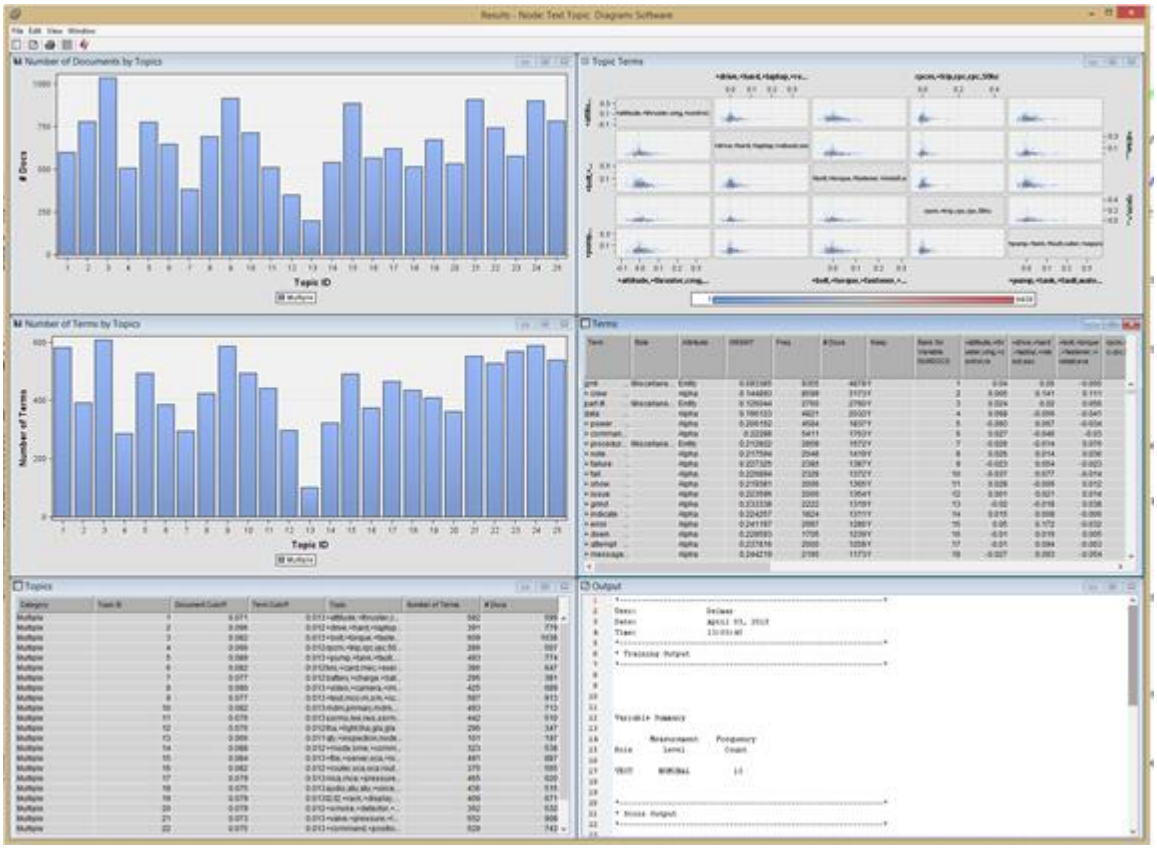


Figure G-8. Text Topic Node Results

Topics

Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
1	0.079	0.005	pcs,sm,+crew,rws,+lab	1738	350
2	0.093	0.004	hardware,usa,+concern,launch pad,1001-1002	48	53
3	0.084	0.004	wis,+remote sensor unit,assy,+wire,+unit	84	52
4	0.077	0.004	red,+indicator,red indicator,+strap,+tether	118	62
5	0.071	0.005	mdm,pl,odin,+dump,mdm	2385	454
6	0.072	0.005	oca,router,oca router,+router,oca	1757	347
7	0.067	0.005	bme,+command,voa,voa,+mode	1828	324
8	0.074	0.005	mec,+card,tvis,+download,cevis	1718	339
9	0.060	0.004	+lot,+tether,+safety tether,load-limiting feature,load-limit...	333	63
10	0.066	0.005	+file,+server,+error,oca,+file	2519	574
11	0.057	0.004	+restraint,+foot,+boot,large,acr	916	80
12	0.096	0.004	+drive,+hard,+hard drive,+laptop,+shell	2149	503
13	0.066	0.005	svs,+test,+video,+display,+failure	2309	296
14	0.060	0.005	+load,+screw,+test,+bracket,+assembly	2881	431
15	0.067	0.005	t2,t2,bme,t2 display,clu	1785	294
16	0.058	0.004	tepc,tepc,tepc mode,mode,ar	1151	130
17	0.063	0.005	+event,mcc-m,+alarm,+pressure,+valve	3251	548
18	0.055	0.004	boeing_hsv,anderson,m,ford,+cell	1819	187
19	0.059	0.005	battery,+battery,emu,+charge,+discharge	2033	254
20	0.065	0.005	rpcm,rpc,rpc,+trip,+power	2437	419
21	0.058	0.005	ssrms,+command,ssrms,+string,+test	2580	347
22	0.069	0.005	pluto,ssc,+network,+connect,ssc	2242	500
23	0.052	0.004	+margin,+gap,pgt,pgt,+insert	904	71
24	0.063	0.005	+card,electronics,+box,+laptop,+network	1363	274
25	0.060	0.005	+download,rsu,iwis,data,+file	2131	285

Figure G-9. Text Topics

SAS® Enterprise Miner and Text Miner produce data files that only SAS® software could read and visualize. Using SAS® tools is not practical from a cost perspective and learning curve standpoint for the entire team, the SAS® Enterprise Guide (EG) tool was used to format the data for Microsoft® Excel® and Tableau®. An EG process flow was constructed to capture the Topic Node data to be used outside the SAS® software (see Figure G-10). Different data tables are appended (hptm_validated and hptm_train), and then a PROC SQL program was run to format the output.

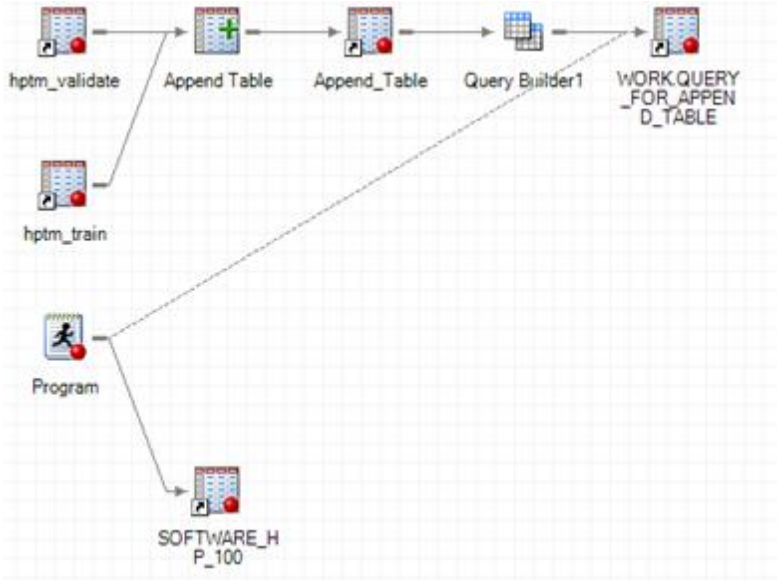


Figure G-10. Enterprise Guide

The Excel® file included the anomaly reports from the merged data set that were associated with the new 25 Topic fields from the SAS® Topic Node output. Approximately 3,875 out of the 13,647 reports were associated with the topics.

To enhance the Excel® file, a Color Coding Application (see Figure G-11) was developed to highlight significant text terms. The records were vetted and color-coded based on the topic and relevant weight. Each term or noun group was color-coded not only by topic term but also by the initial Text Filter Search Expression. The Search Expressions used in the Text Filter node were color coded in blue italic font. The first topic term in each of the topics was blue bold font.

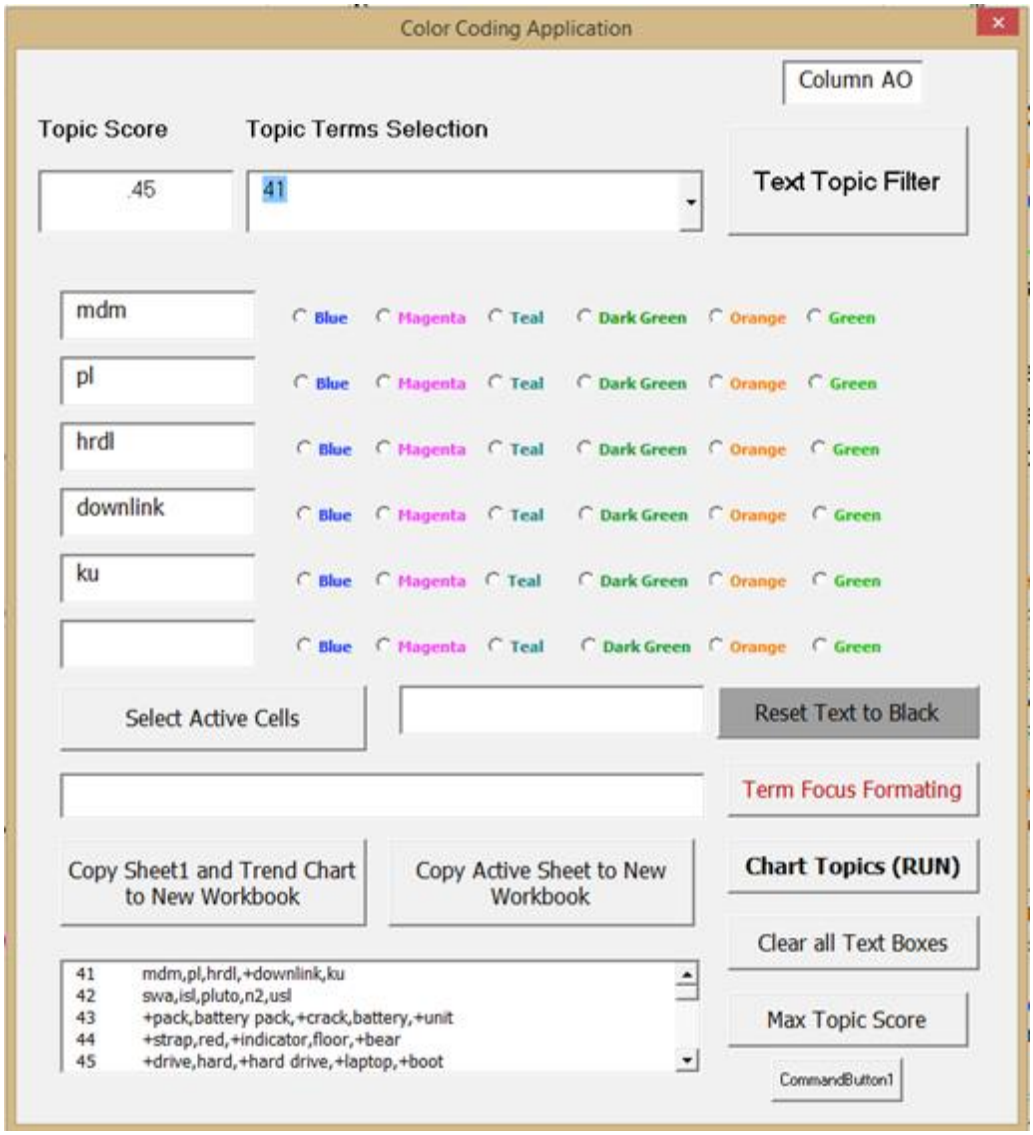
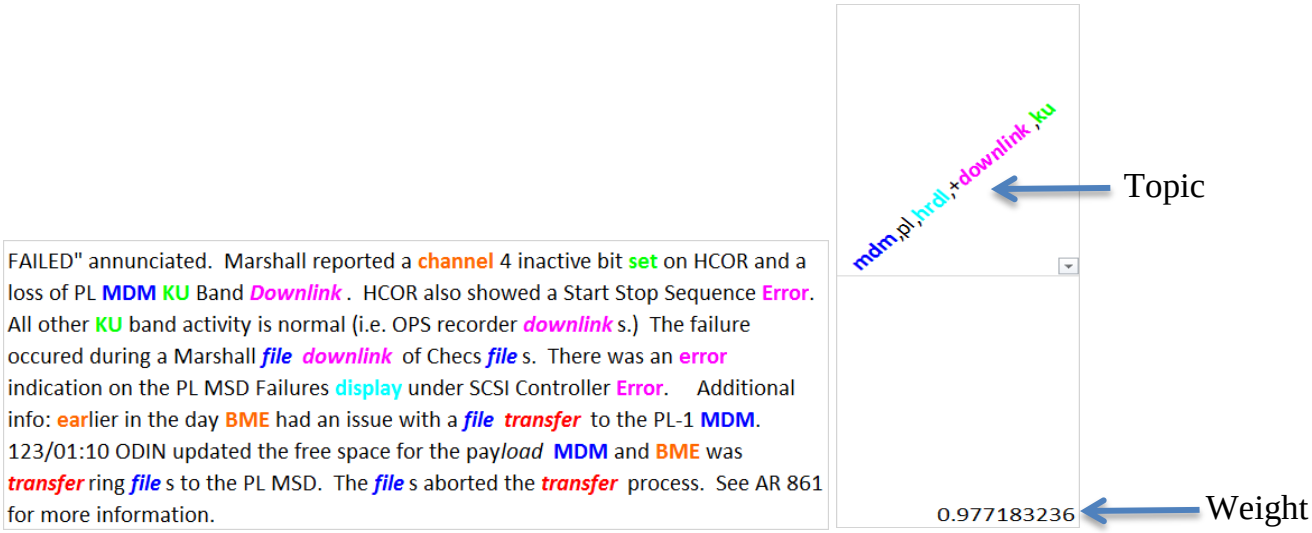


Figure G-11. Excel® Color-Coding Tool

The Problem Description and other textual fields, along with the Topics, were matched and color-coded (see Figure G-12). The relevant weight (see Figure G-12) helped to determine the right level of significance on the specific topic. The goal was to reduce the number of documents to be reviewed without leaving out crucial information. The relevant weight (i.e., a statistical number that was applied by the SAS® software) was filtered based on the manual review (a human review of the data) and by the problem records displayed; the score based on the highest weight.



Detected_Year	Record_Number	Database_Name	TextFilter_relevance	TextFilter_snippet	Problem_Title	Problem_Description	mdm,pl,hrdl,+downlink,ku	swa,sl,pluto,n2,usl	+pack,battery,pack,+crack,battery,+unit	+strap,red,+indicator,floor	+drive,hard
2005	862	MOD AR	0.689304	... during a Marshall file downl	Prim PL HRDL Self Test SCSI TAXI Access	On GMT 2005/123:04:44 ODIN elog "Prim PL HRDL Self Test SCSI TAXI Access Fail is FAILED" annunciated. Marshall reported a channel 4 inactive bit set on HCOR and a loss of PL MDM KU Band Downlink . HCOR also showed a Start Stop Sequence Error. All other KU band activity is normal (i.e. OPS recorder downlink s.) The failure occured during a Marshall file downlink of Checs file s. There was an error indication on the PL MSD Failures display under SCSI Controller Error. Additional info: earlier in the day BME had an issue with a file transfer to the PL-1 MDM. 123/01:10 ODIN updated the free space for the payload MDM and BME was transfer ring file s to the PL MSD. The file s aborted the transfer process. See AR 861 for more information.	0.977183	-0.0025	-0.00645	0.027115	0.064535
2005	1791	PART IFI	0.745638	... during a Marshall file downl	Prim PL HRDL SCSI TAXI Access Fail	On GMT 2005/123:04:44 ODIN elog "Prim PL HRDL Self Test SCSI TAXI Access Fail is FAILED" annunciated. Marshall reported a channel 4 inactive bit set on HCOR and a loss of PL MDM KU Band Downlink . HCOR also showed a Start Stop Sequence Error. All other KU band activity is normal (i.e. OPS recorder downlink s.) The failure occured during a Marshall file downlink of Checs file s. There was an error indication on the PL-1 MDM MSD Failures display under SCSI Controller Error. This condition was seen before on AR 597 and IFI-1638 but was an unexplained anomaly. PRACA 3735 was opened against the PL-2 MDM at the time. This anomaly occurred on PL-1 MDM.	0.976632	-0.00475	-0.00103	0.025174	0.075595
2005	6893	PART PRA	0.586288	... Data Interfaces and Networks</b	Prim PL HRDL SCSI TAXI Access Fail	On GMT 2005/123:04:44 Onboard Data Interfaces and Networks (ODIN) elog "Prim PayLoad (PL) High Rate Data Link (HRDL) Self Test Small Computer System Interface (SCSI) Transparent Asynchronous Transmitter-Receiver Interface (TAXI) Access Fail is FAILED" annunciated. Marshall reported a channel 4 inactive bit set on High Rate Communications Outage Recorder (HCOR) and a loss of PL Multiplexer/Demultiplexer (MDM) KU Band Downlink . HCOR also showed a Start Stop Sequence Error. All other KU band activity is normal (i.e. OPS recorder downlink s.) The failure occured during a Marshall file downlink of Crew Health Care System (CHeCS) file s. There was an error indication on the PL-1 MDM Mass Storage Device (MSD) Failures display under SCSI Controller Error. This condition was seen before on AR 597 and IFI-1638 but was an unexplained anomaly. PRACA 3735 was opened against the PL-2 MDM at the time. This anomaly occurred on PL-1 MDM.	0.974309	0.004665	0.00838	0.023813	0.077326

Figure G-12. Excel® Spreadsheet Color-Coding Examples

The color-coding tool could also be used to create a Tableau® workbook with ITAR banners, trend charts, and sheets for each specific topic. The upper view of Figure G-12 is an exploded view of the lower section. This visualization information was used to determine significant observations or trends that needed further investigation by the discipline expert. Tableau® visualizations are discussed further in Appendix E.

Appendix H. ISS Data Mining Site Construction Guide

ISS Data Mining Site Construction Guide

I. Site Structure and Diagram
- a. Site Diagram ..... page H-3
- b. Modifications ..... page H-4
- c. Page Buttons: HTML, ASP, and Forms ..... page H-4
- d. Workflows ..... page H-4
- e. Site Background Images – CSS link ..... page H-5
- f. Description of Site Settings, Lists, and Libraries ..... page H-6
  - i. ISS Data Mining – Parent Site to Basic Search and Advanced Users page H-6
  - ii. Basic Search Subsite ..... page H-8
  - iii. Advanced Users Subsite ..... page H-10
- g. Dependencies Diagram for Web Part Page Communication at the Advanced Users Subsite ..... page H-11
- h. Basic Search Subsite ..... page H-12

II. Site Features
- a. Save a List or Library as a Template – Description and Awareness of Options ..... page H-12
- b. Versioning on a List or Library ..... page H-13
- c. The Send-To Method ..... page H-13
- d. Site Groups ..... page H-13
- e. Modify/Remove Subsite Links on the Top Links Bar ..... page H-14
- f. The SME view ..... page H-14
- g. Site Modifications for Security and Controlling Access by Groups ..... page H-16
- h. Process for Granting Access to Users ..... page H-17
- i. Site Banners ..... page H-19

Acronyms

SPD – SharePoint Designer. A program used for advanced SharePoint development.

HTML – Hypertext Markup Language. A standardized system for World Wide Web pages.

ASP (ASP.NET) – Active Server Pages. An open source server-side Web application framework designed for Web development to produce dynamic Web pages.

CSS – Cascading Style Sheets. A style sheet language used for describing the look and formatting of a document written in a markup language.

ITAR – International Traffic in Arms Regulations. A set of United States government regulations that control the export and import of defense-related articles and services on the United States Munitions List (USML).

SBU – Sensitive But Unclassified. A designation of information in the United States federal government that, though unclassified, requires strict controls over its distribution.

SME – Subject Matter Expert. An authority in a particular area or topic. The pages on the Basic Search subsite are partitioned according to the SME categories.

XML – Extensible Markup Language. A markup language that defines a set of rules for encoding documents in a format which is both human-readable and machine-readable.

IMCS – Information Management and Communications Support contractor. The team responsible for SharePoint administration and development. The organization to contact for SharePoint technical support.

H.1 Site Structure and Diagram

The Data Mining & Knowledge Management (DM&KM) team site is a SharePoint site collection consisting of a number of subsites accessible through the top link bar. Figure H-1 on the next page shows the site structure. Of particular interest to this report is the ISS Data Mining (ISS DM) subsite due to the ITAR (International Traffic in Arms Regulations) and SBU (Sensitive But Unclassified) nature of the data that it contains. Features were enabled at the site collection level (DM&KM), the top level of the site, in order to enable the functionality required by the ISS Data Mining subsite. Furthermore the ISS Data Mining subsite was modified from the normal SharePoint site structure in order to create a single point of entry to the site at which an Agree-to-Terms popup warning banner must be acknowledged by a visitor in order to proceed. The pages and dialog boxes at ISS Data Mining are also labeled to indicate the sensitivity of the data and to make the user aware of their responsibilities.

While ISS Data Mining is a subsite of the Data Mining & Knowledge Management site, ISS Data Mining is also a parent to two subsites below it. The Basic Search and Advanced Users sites are most directly the subsites of ISS Data Mining and both the Basic Search site and Advanced Users site pull their data from the Site Assets Library of their ISS Data Mining parent.

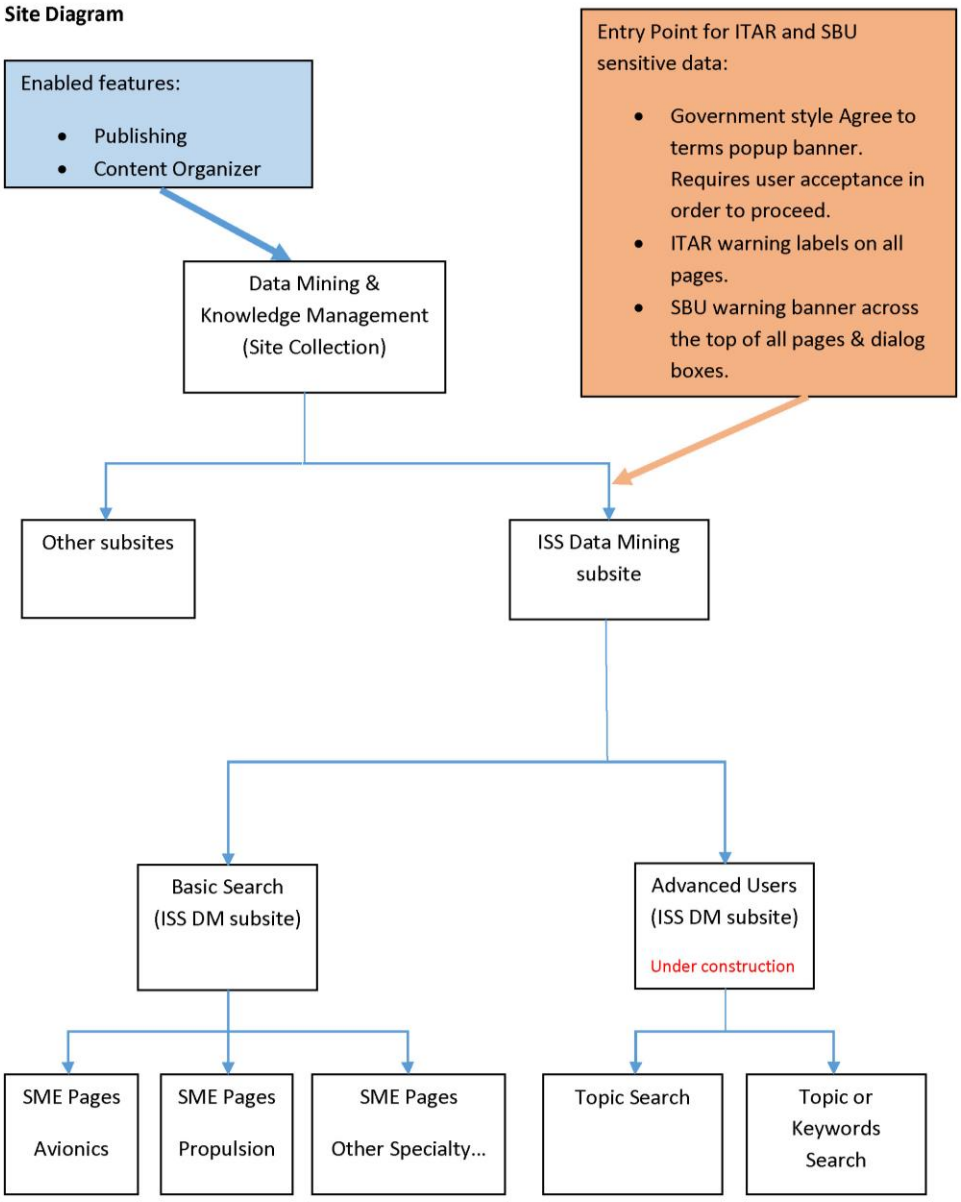


Figure H-1. Diagram of Data Mining and Knowledge Discover Site Collection

Modifications

Note: In accordance with SharePoint best practices, all images and CSS (Cascading Style Sheets) linked resources used in site pages are stored in the Site Assets library. All linked resources for the Basic Search and Advanced Users subsites are stored in the Site Assets library of their parent site, ISS Data Mining.

Note: Links in image and input buttons are not updated on site migrations (moving the test site to the production site) and must be updated manually using SharePoint Designer (SPD) after a migration.

Page Buttons

HTML input, ASP input, and Forms buttons

Site navigation consists of SharePoint link lists and HTML (Hypertext Markup Language) "ab input" buttons. The HTML input buttons are primarily used where a resource has to open in a new window, such as the SME configured windows, and where a list form needs to be activated using the captured properties of the list **add item** control. How a resource is opened is controlled through the use of the window.open() command and its options such as '_self' and 'blank'. The ASP (Active Server Pages) image buttons are used for the link to the Tableau® corporate site and for the information icon placed within the SME (Subject Matter Expert) configured pages found at the ISS Data Mining subsite Basic Search and on the Advanced Users pages. The images used on the ASP image buttons are the official icon for Tableau® Reader, and two concentric circles with an "I" in the middle – created using Paint.net. The window.open() command options used for the information window are 'width=600, height=400, left=600' to control window size, 'blank' to open it in a new window, and 'scrollbars=yes' so that IE, Firefox, and Opera will have scroll bars available in the new window where the site help information appears.

Accessing a list form via an HTML button requires copying the **href** link from the **add item** control usually found at the bottom of a list or library (the small green plus sign). Exposing the href link is done by right-clicking an **add item** button in a list or library then selecting **inspect element** (a browser function). Some **onclick** commands (info icon and form buttons) involve a 'Javascript: return false'; its purpose is to cause the server to process the onclick event but prevent the window from navigating anywhere. Thus, as in the case of the information button, a new window with site help information will appear but the page itself where the button was clicked will not change.

Workflows

There are two types of workflows used on the site. List workflows that cannot be moved with the site upon a migration and reusable workflows which can move with the site.

Site Background Image – A CSS link to the Source Image in the Site Assets Library

Figure H-2. Site Background Image

Image link: <link rel="stylesheet" type="text/css" href="https://spstest.ksc.nasa.gov/sites/dm/ISS-DM/SiteAssets/Kevin_S_Ribbon_TopBar_Body_Custom_CSS.css"/>

Two background images are currently used on the site:

- At the Basic Search default page and subdisciplines page: Background_ISS_large_210.jpg
- At the Advanced Users default page: Background_ISS_large_210_UC.jpg. This is the same image but with the words "The Advanced Site is Currently" "Under Construction" centered, with the latter half beneath the former.

Both images, like **all images and page resources**, are stored in the **Site Assets Library** of the ISS Data Mining site. ISS Data Mining serves as the central repository for all data that is common between the Basic Search and Advanced Users subsites.

Description of Site Settings, Lists, and Libraries

ISS Data Mining – Parent Site to Basic Search and Advanced Users

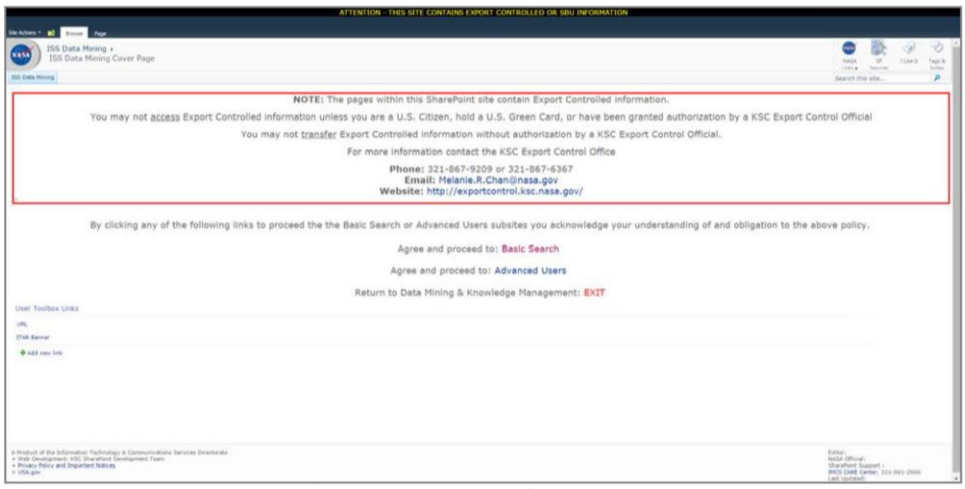


Figure H-3. ISS Data Mining Cover Page

With Content Organizer rules and Publishing enabled on the site, the ISS Data Mining site settings page resembles Figure H-4.

Figure H-4. ISS Data Mining Site Settings Page

The currently enabled site features under the **Site Actions → Manage site features** are:

- Content Organizer

- Metadata Navigation and Filtering
- NASA Site Properties
- Offline Synchronization for External Lists
- Team Collaboration Lists

Figure H-5 shows the lists and libraries that are affected by workflows as described in each section below. They are also affected by the Content Organizer rule that any file with the ".twbx" extension uploaded to the site will be routed to the **Visualization files Library**. All other files will be automatically routed to the **Drop Off Library**.

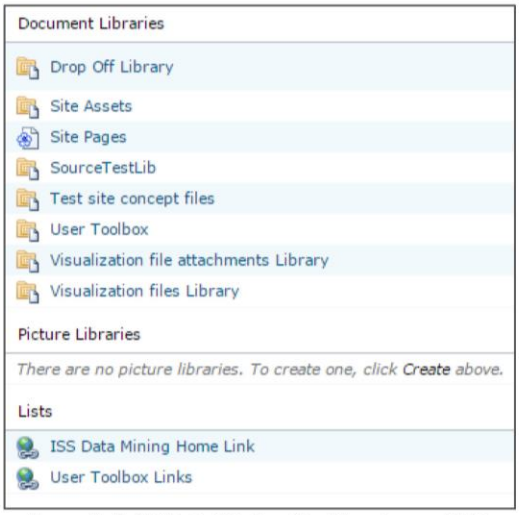


Figure H-5. ISS Data Mining Site Libraries and Lists

Drop Off Library – automatically created by SharePoint when the Content Organizer feature is enabled.

Site Assets – This is where you will find the elements of the site help pages, the CSS script that attaches the ISS background image to some of the site pages, the images used for the information icon, the Tableau® icon, and other site page elements. Other examples are the white text and black text ITAR banners stored in Site Assets and referenced by an **XML Viewer Web Part** embedded into each page that displays those banners. The XML (Extensible Markup Language) pages for the ITAR banners were created in SPD (SharePoint Designer) and use standard XML coding (some of the coding used in our SharePoint pages is not recognized by a standard XML processor).

User Toolbox – stores tools for site users. Currently holds image copies of the ITAR/SBU Agree-to-Terms banner and Export Control banner.

Visualization file attachments Library – for file attachments to be associated with particular visualization files. When entering a new item, the properties must be filled out, one of which will be what visualization file to associate to the newly entered attachment.

Visualization files Library – this is the library for storing the Tableau® visualization files that the subject matter experts (SMEs) and other site members will have access to. The library at the ISS Data Mining site is considered the source and the parent for the same libraries at the Basic Search and Advanced Users subsites. Visualization files are stored in the library at the ISS Data Mining site and then virtual copies are sent to the subsites using the **send-to** functionality.

ISS Data Mining Home Link and the User Toolbox Links - The first is a link to the ISS Data Mining site (used in some pages) and the second provides links to images and resources useful to a user.

Other libraries and lists may exist that were not created for the ISS Data Mining site project and may be the product of other site administrators such as IMCS technicians.

Basic Search and Advanced Users – Subsites below ISS Data Mining

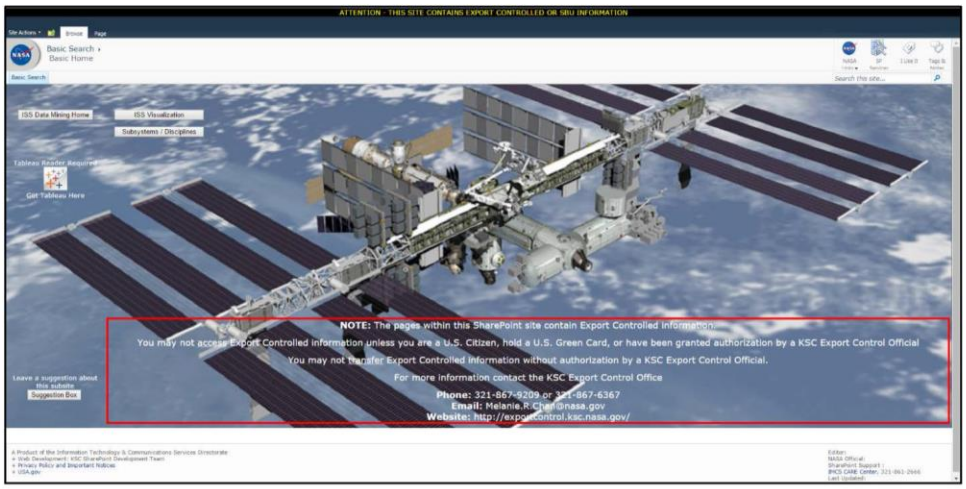


Figure H-6, Above. Basic Search Home Page

Figure H-7, Left. Advanced Users Home Page

The site settings page for both Basic Search and Advanced Users should be inherited from the parent site, ISS Data Mining, and should look the same as on page H-6.

The lists and libraries in the subsites are the same as on the parent. The subsites Basic Search and Advanced Users will both have the same information as the parent site, ISS Data Mining, but will each contain additional lists and libraries as needed for their individual functionality.

The currently enabled site features for both subsites, Basic Search and Advanced Users, under the **Site Actions → Manage site features** are:

- Content Organizer
- Metadata Navigation and Filtering
- Offline Synchronization for External Lists
- Team Collaboration Lists

Both subsites have a **Suggestion Box** list and the **Request new or modified visualization** list. For users to leave suggestions and request new or modified visualizations respectively.

Basic Search libraries and lists

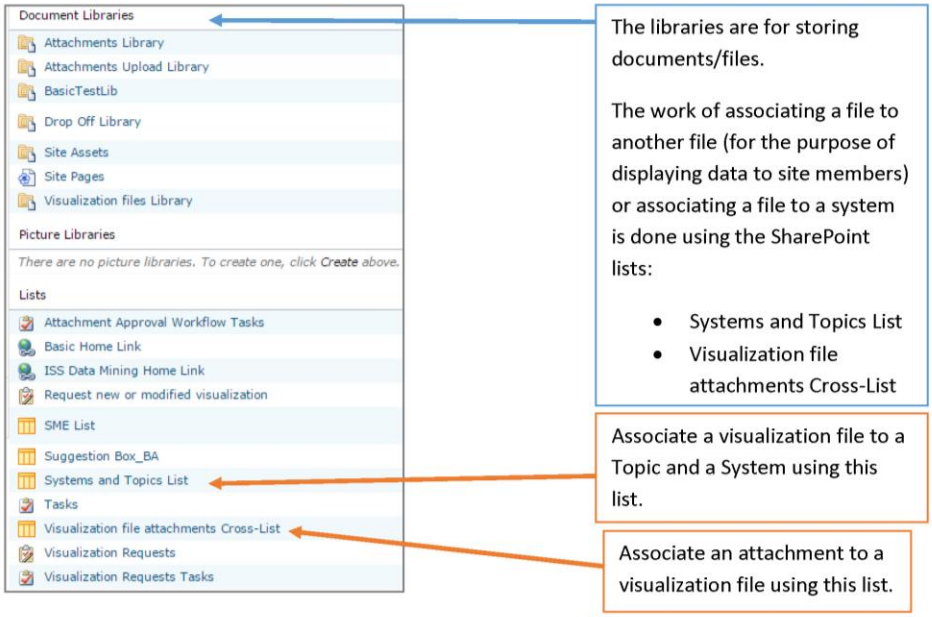


Figure H-8. Basic Search Libraries and Lists Overview

Libraries

All lists and libraries have the same purpose as the ones in the parent site, ISS Data Mining (see page H-7, Figure H-5), except as listed below.

Site Assets – this library is intended to be unused in both subsites below ISS Data Mining since both subsites require the same assets and therefore should access them from the parent site.

Lists

SME List - the Subject Matter Experts list. This list is the basis for the system/discipline categories available in the Systems and Topics List and the basis of the partitioning on the Basic Search subsite.

Systems and Topics list – a list of systems and their associated topics and visualization files.

Suggestion Box – has versioning activated so users can add to the **Comments/Suggestions** field and append new comments to their old ones.

Visualization file attachments Cross-List – the list for associating an attachment to a visualization file.

Visualization Requests List & Request a new or modified visualization List – these lists are meant to have workflows attached to them for the purpose of handling site member requests for visualizations. Once a reusable type workflow for visualization requests is complete then it will be attached to the Visualization Requests List and the other list will be deleted.

Advanced Users – Subsite below ISS Data Mining

The Advanced Users site is not currently a project priority and so is marked as "Under Construction" as shown in Figure H-7 on page H-8. The lists, libraries, and workflows on this site may lag behind the higher priority Basic Search site in terms of features and functionality. This site has two pages for advanced searches. One is intended for content search by topic, the other by topic or word cluster. Both pages exist and the only missing functionality is the search by word cluster.

The lists and libraries on the Advanced Users site are the same as at Basic Search except as noted below.

Libraries

Attachments Library & Attachments Upload Library – these libraries were intended to be used in conjunction with an approval workflow to allow site members to submit files to be considered for attachment to visualizations.

ListsTopics and Clusters List – This list relates a topic to a cluster, or group of words. This purpose of this list comes from SAS generating topics and the words associated with them. The idea is to help SMEs by allowing them to type search words which would then expose the topics associated with those words. Those topics are then related to systems and visualization files via the System and Topics list.

Advanced Users Links, ISS Data Mining Home Link, Topics Search I and II Links, Link to Topic or Keywords Search – these are all lists containing a link to the target as named in the list title. These lists are used in Data View web parts embedded in some pages allowing the user to click and change pages to the target that is indicated in the name of the list.

Dependencies Diagram – Web Part Communications at the Advanced Users Subsite

Structure of Web Part Communication, Topic Search Navigation Page, and Topic Search Page

The following diagram is a guide to setting up the communication between web parts when using SharePoint Designer 2010. The Data View web parts use a producer/consumer model of sending data from one web part to the other. Each consumer uses data from the producer (or upline web part) in order to filter the view that it displays on the Web Part Page. This Diagram also indicates the fields within a list or library that must be completed in order for new items to properly link and display on a page view.

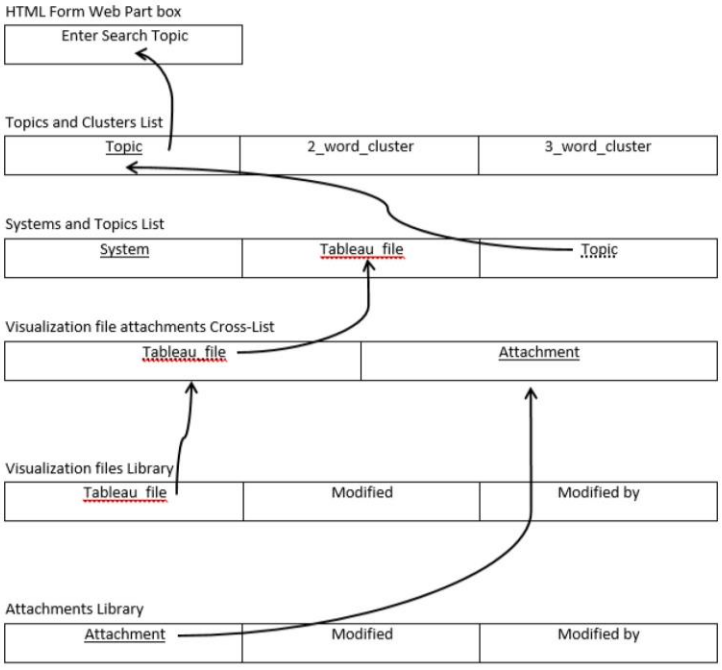


Diagram of Web part communication (arrows are in the direction of consumer to producer)

Figure H-9. Structure of Web Part Communication for SME Pages, Advanced Users

Basic Search Subsite

The following diagram is a guide to setting up the communication between web parts when using SharePoint Designer 2010. The Data View web parts use a producer/consumer model of sending data from one web part to the other. Each consumer uses data from the producer (or upline web part) in order to filter the view that it displays on the Web Part Page. This Diagram also indicates the fields within a list or library that must be completed in order for new items to properly link and display on a page view.

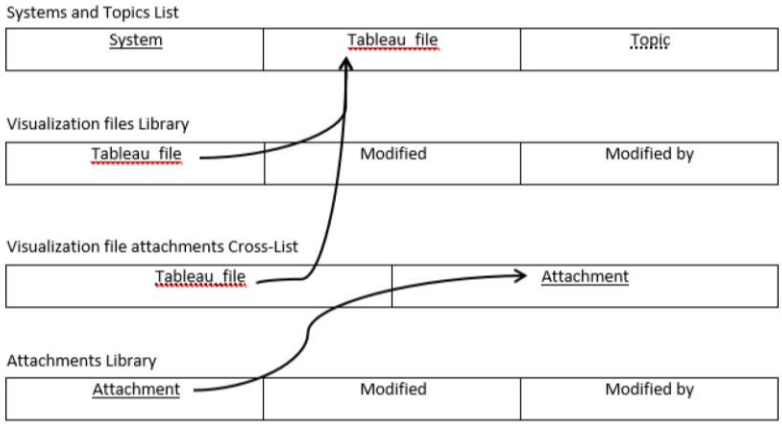


Diagram of Web part communication (arrows are in the direction of consumer to producer)

Figure H-10. Structure of Web Part Communication for SME Pages, Basic Search

H-2. Site Features

Save List or Library Content as a Template – Awareness of Options

Avoid reentry of data for lists such as the Systems and Topics list by saving it as a template. After the template is created, it can be downloaded as a file with a SharePoint Template File (.stp) extension and uploaded back to SharePoint via the **List Template Gallery** at Data Mining & Knowledge Management; the top site in the collection. Creating a template is a useful means of backing up a list or library and a necessary intermediate step to downloading a copy for backup to other media outside of SharePoint.

For security reasons, data such as that in the Visualization files Library which contains ITAR (International Traffic in Arms Requirements) and SBU (Sensitive But Unclassified) data should **not** be saved as a template *along with its content* as this would allow sensitive data to be recovered and viewed outside of the security of the site.

Versioning

Required in the suggestion box list in the **Multiple Lines of Text** fields when the option to append comments is selected in "column type" setup.

Settings:

- Require content approval for submitted items?
  - No
- Create a version each time you edit a file in this document library?
  - Create major versions
- Optionally limit the number of versions to retain:
  - Check the box to **Keep the following number of major versions:**
  - 5
- Who should see draft items in the document library?
  - Any user who can read items
- Require documents to be checked out before they can be edited?
  - Yes

Send-To: Creating Virtual Copies of a File in Other Locations

Useful in the ISS Data Mining subsites because both Basic Search and Advanced Users subsites require exactly the same files. We keep a single copy of a visualization file in the **Visualization files Library** at the ISS Data Mining site and use the **send-to** functionality (requires Publishing to be enabled on the site) to send virtual copies to the file libraries at the Basic and Advanced subsites. Update files by following the proper check-in/check-out procedure.

Site Groups

Two groups created for the project are:

- Data Mining Approvers – a group used by the Attachment Approval workflow for approving attachments to visualization files. The Attachment Approval workflow is a reusable type workflow and is still a work in progress.
- Data Mining Resolvers – a group used by the Visualization Request workflow to assign visualization request tasks. This group should be emailed whenever a site member creates a new visualization request. The Visualization Request workflow is a list type workflow that will eventually be converted into a reusable type workflow.

Modify/Remove Subsite Link on the Links Bar

We do not want site members to be able to access the subsites of ISS Data Mining without clicking on the specific subsite links that we created at the bottom of the site cover page. So the subsite links were removed from the top links bar and only the ISS Data Mining subsite serving as the parent to subsites beneath it should be made to show on the top links bar.

Note: It was necessary to change the subsite display settings, revert back to normal, then change the settings again in order to remove unwanted subsites from the top links bar.

Once the procedure is completed, the top links bar should only show the ISS Data Mining subsite and not the links to Basic Search or Advanced Users as in Figure H-11 below.

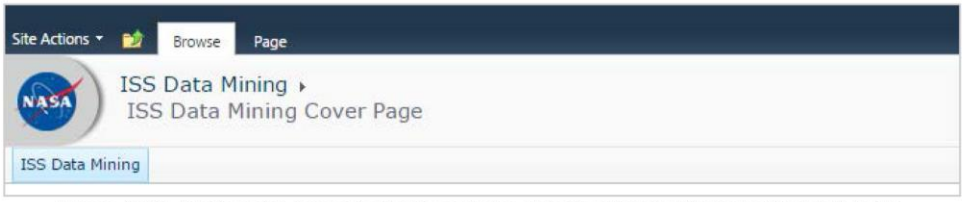


Figure H-11. Display of a Properly Configured Top Links Bar for the ISS Data Mining Subsite

The SME View

The SME view is a Web Part page based on the Header, Left Column, Body format found in either SharePoint Designer or in SharePoint: Site Actions → More Options → Page → Web Part Page. Figure H-12 on the next page shows the format and the page built upon that format.

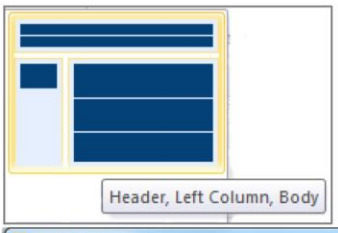
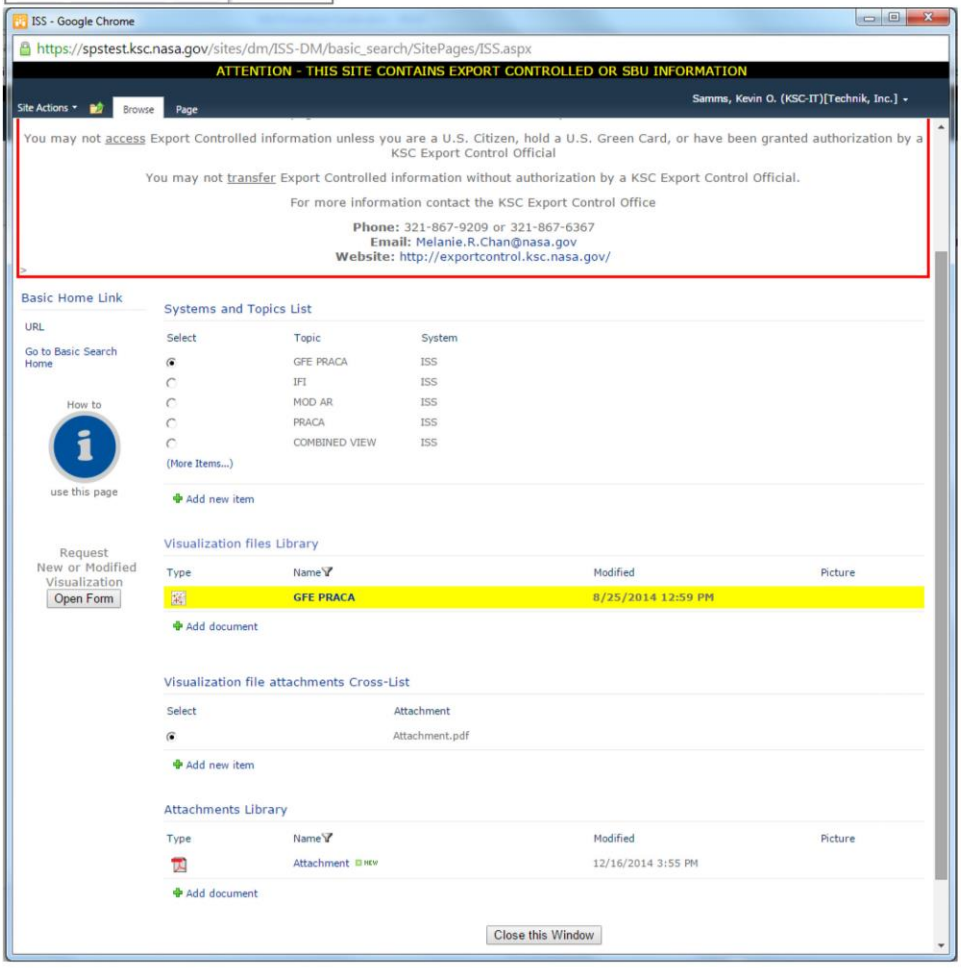


Figure H-12, Left. SME Page Format

Figure H-13, Below. SME Built Page



The Body of the SME window consists of four Data View 2eb parts, each connected to a list or library. Each web part communicates with other web parts following the diagram of web part communication on pages H-12 and H-13. The list and library connection for each Data View web part is shown below.

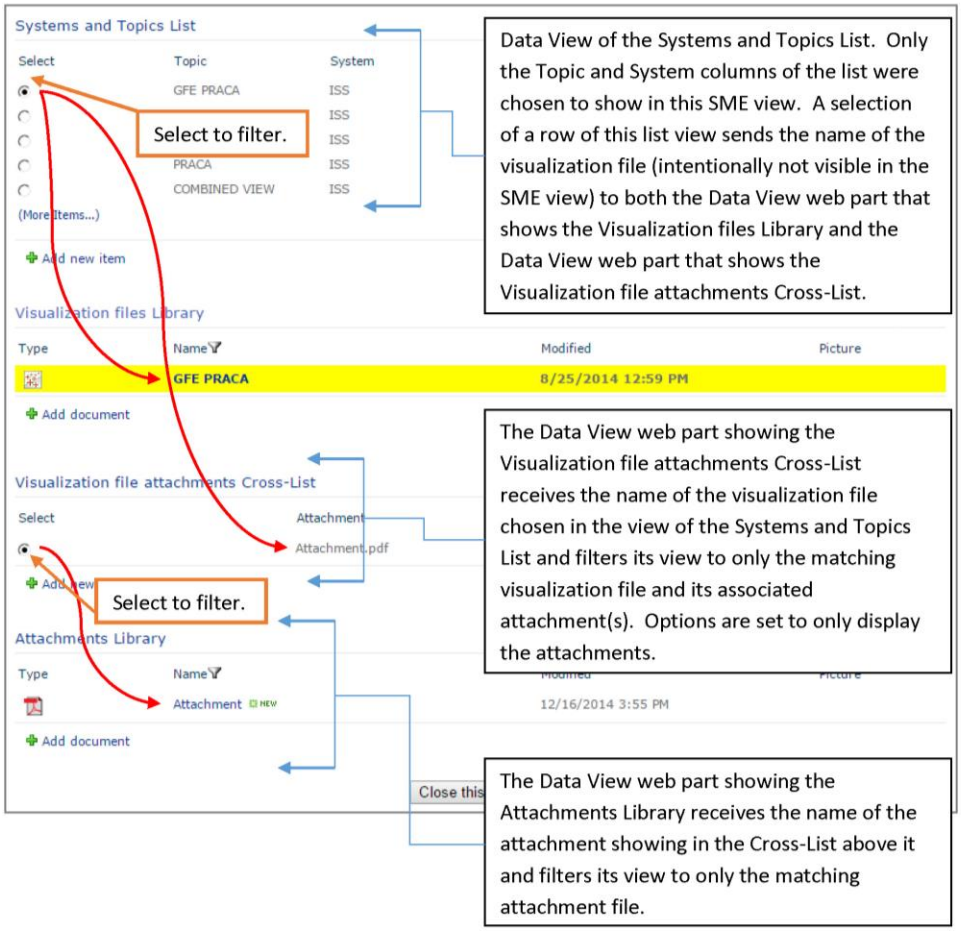


Figure H-14. Diagram of Library and List Communication Following Figure H-10, Page H-13

Update the SME view by adding the appropriate entries to the lists which control library filtering.

Site Modifications for Security and Controlling Access by Groups

To control access to the ISS Data Mining subsite (ISS DM), and consequently the subsites Basic Search and Advanced Users below it, inheritance from the parent site Data Mining and Knowledge Management (DM&KM) was stopped and the parent site group permissions were removed from the ISS DM subsite. Security groups were then created within the ISS DM subsite in order to isolate ISS DM members from all other sites within the DM&KM site collection. The combination of broken inheritance and having subsite specific security groups allows the ISS DM subsite to be unaffected by changes to security groups at the parent site. Thus users added to groups at the parent, DM&KM, are not

automatically granted access to the ITAR/SBU sensitive ISS DM subsite. Likewise, members added to the ISS DM groups have no access to the parent site, DM&KM, or any of its *other* subsites. The standard SharePoint groups; Owners, Designers, Members, and Visitors were created under the ISS Data Mining name at the ISS Data Mining subsite to maintain the SharePoint security group naming convention. To complete the modification and isolation of the ISS DM groups, members were moved from their DM&KM groups to their respective groups at ISS DM.

For example, users in the **ISS Data Mining Members** group have permissions to the ISS Data Mining subsite, its ITAR/SBU sensitive data, and the two sites below it but will not see anything in the DM&KM parent site or any of its other subsites – unless they are separately given permissions to a security group belonging to the parent site, such as the **Data Mining & Knowledge Management Members** group. Likewise any users granted permissions to any parent site group such as the **Data Mining & Knowledge Management Members** group will not be able to see the ISS DM subsite – unless they are separately also placed into an ISS Data Mining group.

Process for Granting User Access to ISS Data Mining

A potential ISS Data Mining site user must first obtain access credentials, on their own, to the following systems.

Government Furnished Equipment Problem Reporting and Corrective Action (GFE PRACA)

ISS Maintenance and Analysis Data Set (MADS)

ISS PART

Quality Assurance Record Centers (QARC) Web Reports

ISS Data Mining external users are **not** granted access to other secured sites.

Once the proper credentials are obtained, the user will contact one of the site administrators listed below:

Ali Shaykhian	Delmar Foster
ali.shaykhian@nasa.gov	delmar.c.foster@dataminingusa.com
321.861.2336	321-867-6631

The site administrator will use IdMAX to verify that the user has obtained the proper credentials and then add the user to the group container **ISS Data Mining Members** or **ISS Data Mining Visitors** of the ISS Data Mining site as appropriate. The Members group receives **Contribute** permissions and the Visitors group receives **Read** permissions.

Agree-to-Terms Popup Banner

Upon first visit to the site a user will see the popup banner shown in Figure H-15. Clicking OK on the banner stores the ID of the user and the time that they accepted the terms. This information according to IMCS. The user will see the banner again when the cached user credentials have expired.

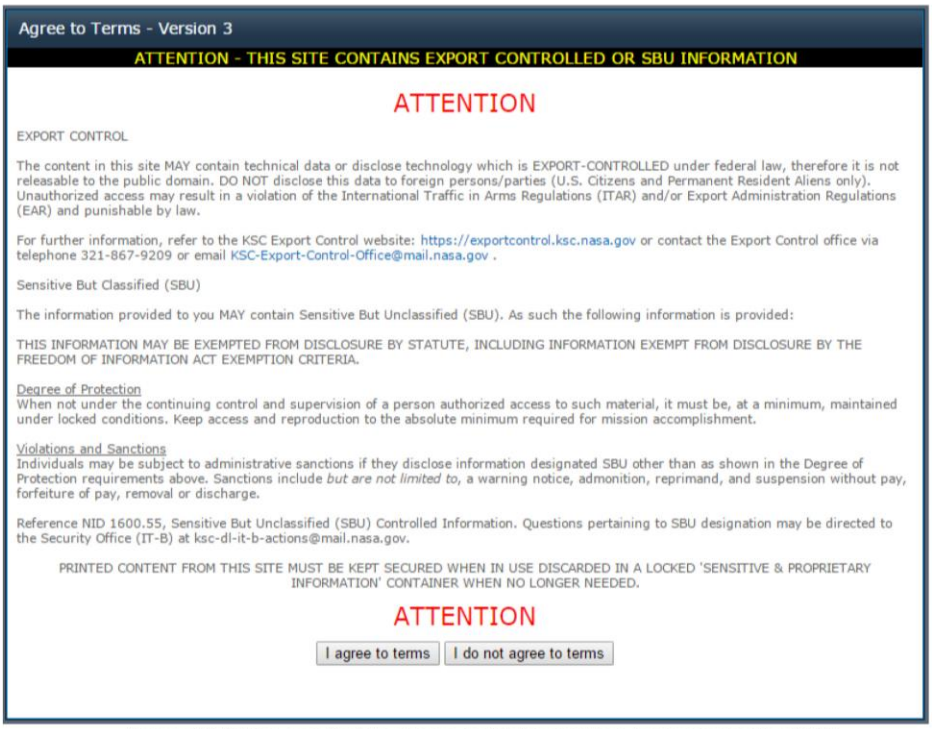


Figure H-15. The result of the Agree-to-Terms markup implemented by IMCS

Label for all Pages and Export Controlled Content

All pages and site content, such as Excel files and Tableau® workbooks, are be labeled with the following banner across their uppermost visible area.

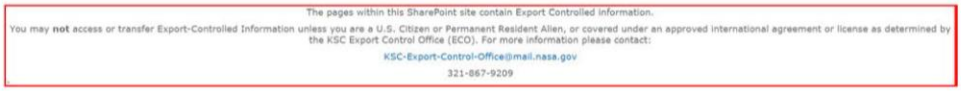


Figure H-16. The result of the markup stored as XML in the Site Assets Library and embedded into an XML Viewer Web Part

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01- 12 - 2015	Technical Memorandum	May 2014 - September 2015

4. TITLE AND SUBTITLE
International Space Station (ISS) Anomalies Trending Study
Appendices

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Beil, Robert J.; Brady, Timothy K.; Foster, Delmar C.; Graber, Robert R.; Malin, Jane T.; Thornesbery, Carroll G.; Throop, David R.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
869021.01.07.01.01

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NASA Langley Research Center
Hampton, VA 23681-2199

8. PERFORMING ORGANIZATION REPORT NUMBER
L-20649 NESC-RP-14-00950

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSOR/MONITOR'S ACRONYM(S)
NASA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
NASA/TM-2015-218991/Volume II

12. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified - Unlimited
Subject Category 59 Mathematical and Computer Sciences (GENERAL)
Availability: NASA STI Program (757) 864-9658

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The NASA Engineering and Safety Center (NESC) set out to utilize data mining and trending techniques to review the anomaly history of the International Space Station (ISS) and provide tools for discipline experts not involved with the ISS Program to search anomaly data to aid in identification of areas that may warrant further investigation. Additionally, the assessment team aimed to develop an approach and skillset for integrating data sets, with the intent of providing an enriched data set for discipline experts to investigate that is easier to navigate, particularly in light of ISS aging and the plan to extend its life into the late 2020s. This document contains the Appendices to the Volume I report.

15. SUBJECT TERMS
Data Mining; Anomaly; International Space Station; NASA Engineering and Safety Center

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			STI Help Desk (email: help@sti.nasa.gov)
					19b. TELEPHONE NUMBER *(Include area code)*
U	U	U	UU	115	(443) 757-5802