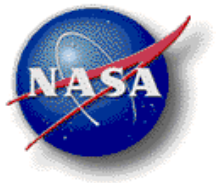
Semantic Annotation of Complex Text Structures in Problem Reports

Jane T. Malin, David R. Throop and Land D. Fleming

Text analysis is important for effective information retrieval from databases where the critical information is embedded in text fields. Aerospace safety depends on effective retrieval of relevant and related problem reports for the purpose of trend analysis. The complex text syntax in problem descriptions has limited statistical text mining of problem reports. The presentation describes an intelligent tagging approach that applies syntactic and then semantic analysis to overcome this problem. The tags identify types of problems and equipment that are embedded in the text descriptions. The power of these tags is illustrated in a faceted searching and browsing interface for problem report trending that combines automatically generated tags with database code fields and temporal information.



Semantic Annotation of Complex Text Structures in Problem Reports

Jane T. Malin, David R. Throop and Land D. Fleming
NASA Johnson Space Center, The Boeing Company and MEI

NASA Conference on Intelligent Data Understanding
Mountain View, California
October 19, 2011

Overview

- Information Extraction for Problem Reports
- Advanced syntactic analysis
- Ontology-based semantic annotation
- User Interface for Analysts
- Evaluation
- Conclusion and What's New

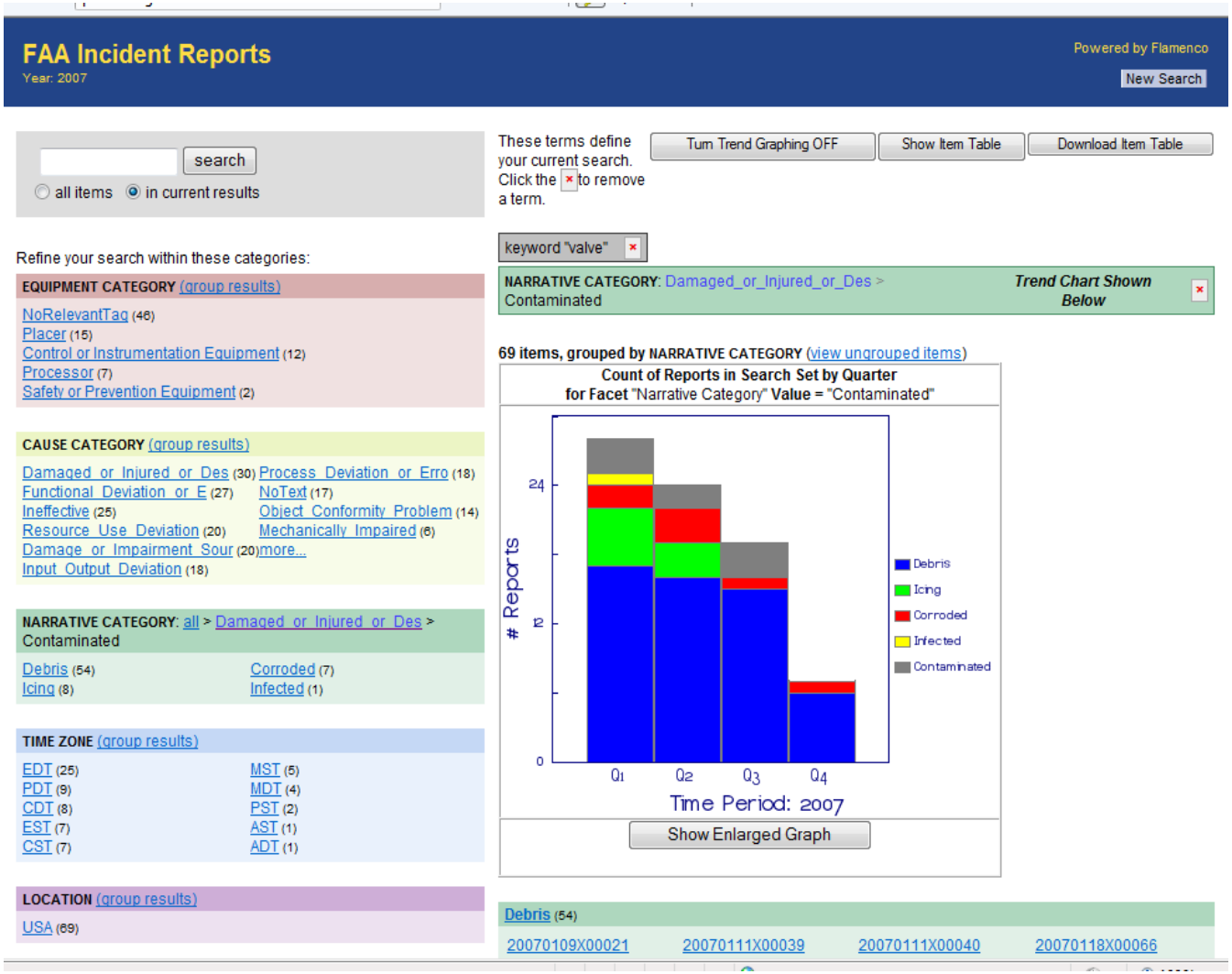
Problem Report Analysis

- Analysts find groups of similar problems to:
  - Identify causes and corrective actions with wide impacts, and look for time trends
  - Get ideas on handling a new problem or mishap by identifying similar past problems
- Using text descriptions to find similar problems
  - Search using codes and keywords is ineffective
    - Misleading codes, false alarms and misses, complex queries
  - Statistical text mining is hard to interpret
    - Identifies groups but is not usually guided by search goals
    - Ignores complex syntax - false alarms and misses
  - Linguistic analysis can be specialized and complex

Information Extraction

- Goal: Automatically extract structured information from unstructured and semi-structured text fields that describe problems
- Linguistic Approach: **Semantic Text Analysis Tool (STAT)**
  - First use practical and general syntactic analysis
    - Specialized training sets not required
    - **Minimal Clausal Reconstruction (MCR)** algorithm from Dr. F. Gomez, University of Central Florida
    - Builds on results from Stanford/Charniak parser
  - Next use hierarchy of types in lexicalized **Aerospace Ontology (AO)** for semantic analysis
  - Associate the semantic annotations (tags) with the data records that contain the text fields
  - Improve search, browsing and data mining

Flamenco+ Web Display



- Data fields tagged with equipment types & problem types
- Search for Valve
- Browse to Damaged…> Contaminated
- Automatic bar graph shows trends
- Buttons to show or download table

Match a Concept with its Modifier

- Goal: Identify and tag types of problems in problem description text fields such as those in Discrepancy Reports (DRs)
  - Text associates bad properties (discrepancy modifiers) with concepts (objects, occurrences or properties)
- Challenge: Modifiers are frequently separated from concepts in natural language problem descriptions
  - Intervening dependent clauses or negations (for example, *machine screw located on… panel is not seated correctly*)
  - Intervening conjuncts (for example, *the door had inadequate paint and good clearance)*
  - The concept is not the head of a noun-phrase (for example, *it passed the insufficient-clearance test)*

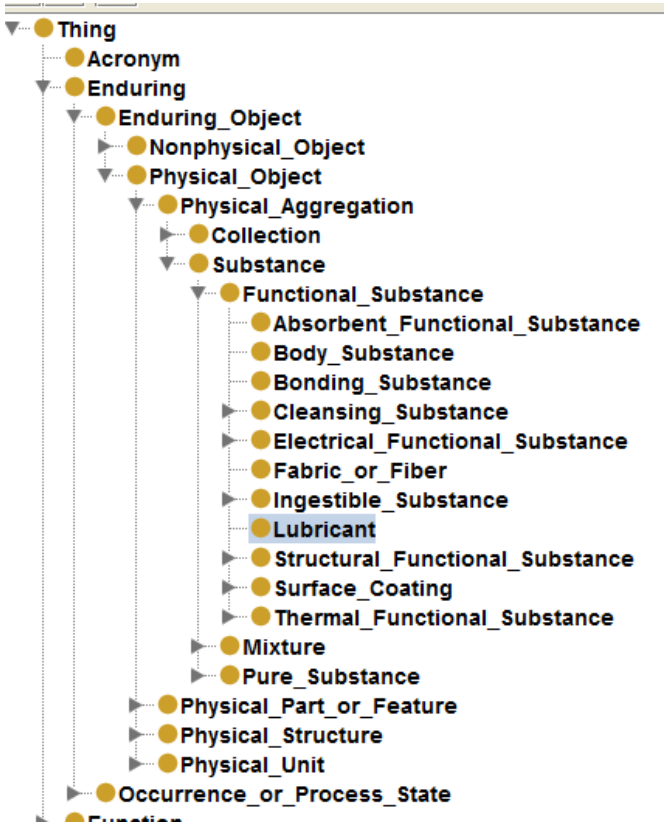
Syntactic Clausal Reconstruction

- Solution: Use MCR syntactic clausal reconstruction algorithm to match the modifiers with the right concepts
  - Resolves empty nodes in parse trees
  - Uses syntactic rules to determine complements and adjuncts
  - Resolves the clause structure for each verb (argument + adjuncts), determining all clausal modifiers for each verb

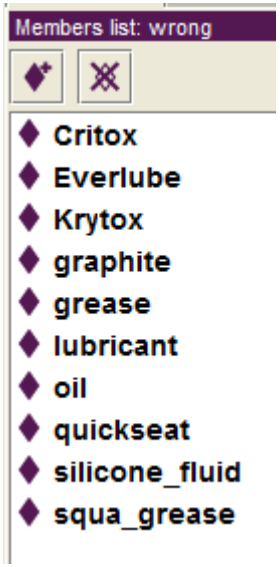
Ontology for Semantic Analysis

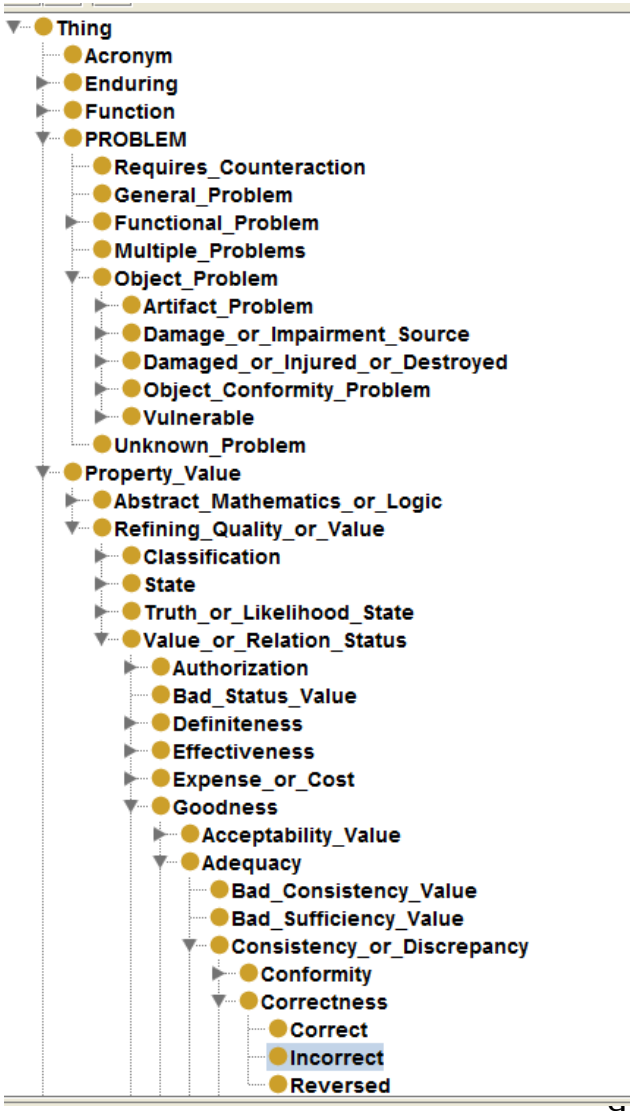
- Lexicalized ontology
  - Each concept is extended with a list of words or phrases that are possible text representations of the concept
- Properties ontology for use in problem description
  - Good and bad
- Phrases in lexical lists in the ontology capture contextual distinctions
  - Warm beer vs. cold coffee

Aerospace Ontology



Lubricant





Lexical Problem Phrases

- Problem phrases combine types of
  - Negative properties and values
  - Objects, occurrences, actions and functions
- Phrases in a lexical list can be defined as combinations of terms from other categories
- Example:
  - Lubricant_Problem: (Excessive, Insufficient, Incorrect, Missing) (Lubricant)
  - (Incorrect) (Lubricant) expands to “wrong lubricant,” “improper Everlube,” “incorrect grease,” and many more

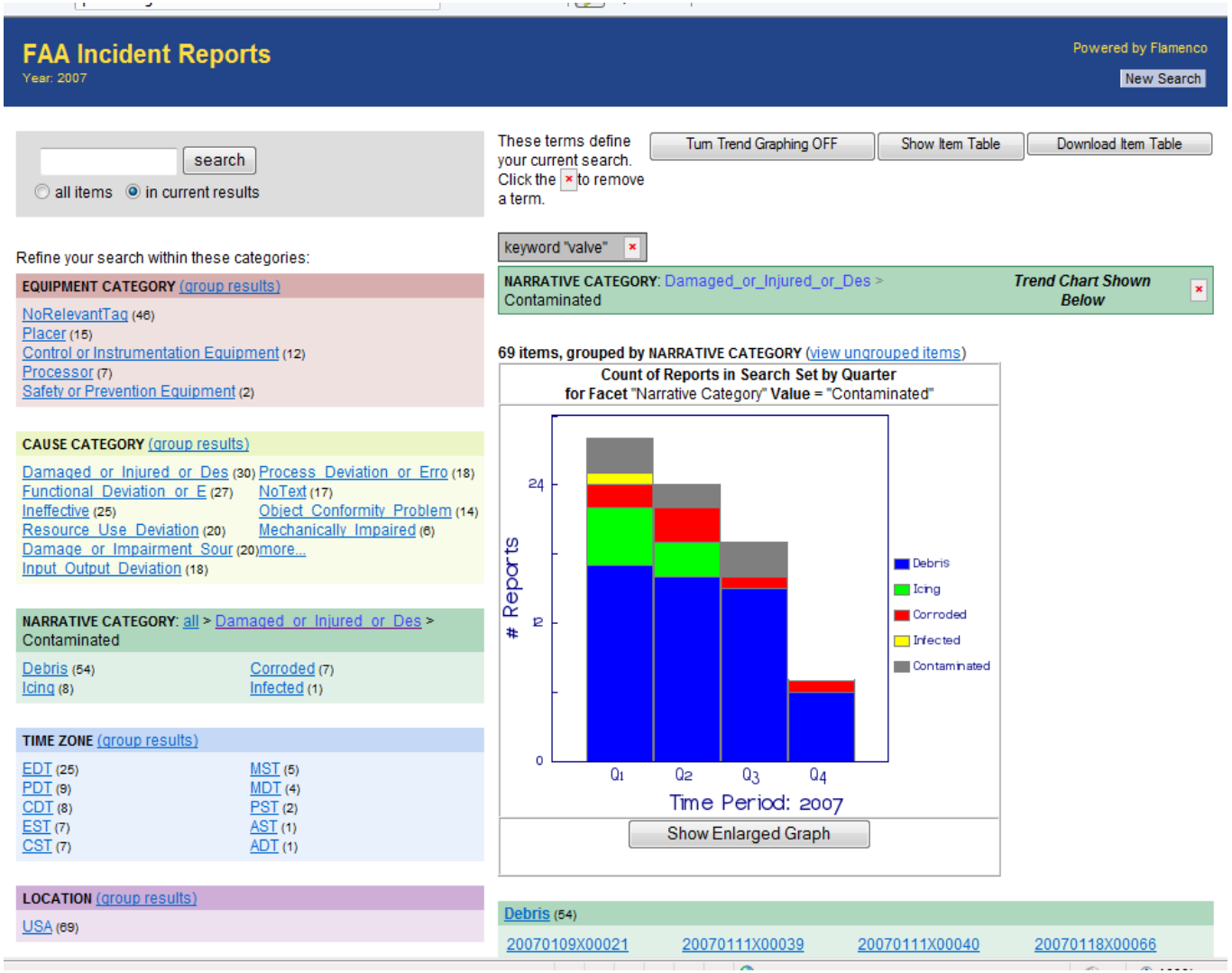
STAT Linguistic Analysis

- Lemmatize words to canonical forms
  - Stem words and phrases in MCR clauses and the problem type hierarchy in AO
- Match
  - Phrases (concepts and their modifiers) in the MCR clauses
  - Terms in the lexical lists in AO
- Use matches to assign problem tags to data records
  - Problem types in AO problem hierarchy

Hierarchical Tags

- Information extraction enables more effective grouping of DRs by problem type
- Extracted hierarchical tags, codes and original text can be used in combination
  - Graph time trends
  - Search and filter in a hierarchical browser
  - Mine the data records with the added tags

Flamenco+ Web Display



- Data fields tagged with equipment types & problem types
- Search for Valve
- Browse to Damaged…> Contaminated
- Automatic bar graph shows trends
- Buttons to show or download table

Clausal Reconstruction for STAT Accuracy

- Using Clausal Reconstruction substantially improves tagging accuracy for problem reports
- Method
  - 36 problem categories from 2007-2008 set, sample of 200 DRs
  - Manual scoring: 101/200 DRs matched at least one category
  - Measures of Accuracy with MCR algorithm
    - Recall: proportion of all true cases tagged (87/101 = 0.86)
    - Precision: proportion of tagged cases that are correct (87/111 = 0.78)

STAT tagging method	Recall	Precision	True positive	False negative	False positive
STAT without MCR algorithm	0.10	0.27	10	91	27
STAT with MCR algorithm	0.86	0.78	87	14	24

Improving Text Mining with Tags

- Analyst goal is to increase Recall
  - Improve search by finding more true positives
- Text miner: Quantum Text
  - Search results are used to define exemplars (5 positive examples for training), for better retrieval
  - Result lists are ranked by similarity
- Compare Search, Text Miner, Text Miner with STAT Tags
- Evaluation Method
  - 2,000 DR records from FY2008
  - 10 test cases with 9-41 true records each (# true records = 249), selected from 36 problem types used for the 1st study
  - Score retrieved records for each case: 1.5 x # true records found by search, and not the 5 exemplars (# true records= 199)

Improvements in Text Mining

- True positive (T_p), false negative (F_n) and false positive (F_p) columns show total frequencies for all 10 cases.
- Text mining retrieval results without STAT tags are disappointing - 38% average recall [T_p / (T_p + F_n)]
  - Consistent with low recall (21%) in MedScan text mining
  - Substantial increase in F_p reduced average precision to 37%
- Using STAT tags substantially improved text mining Recall and Precision

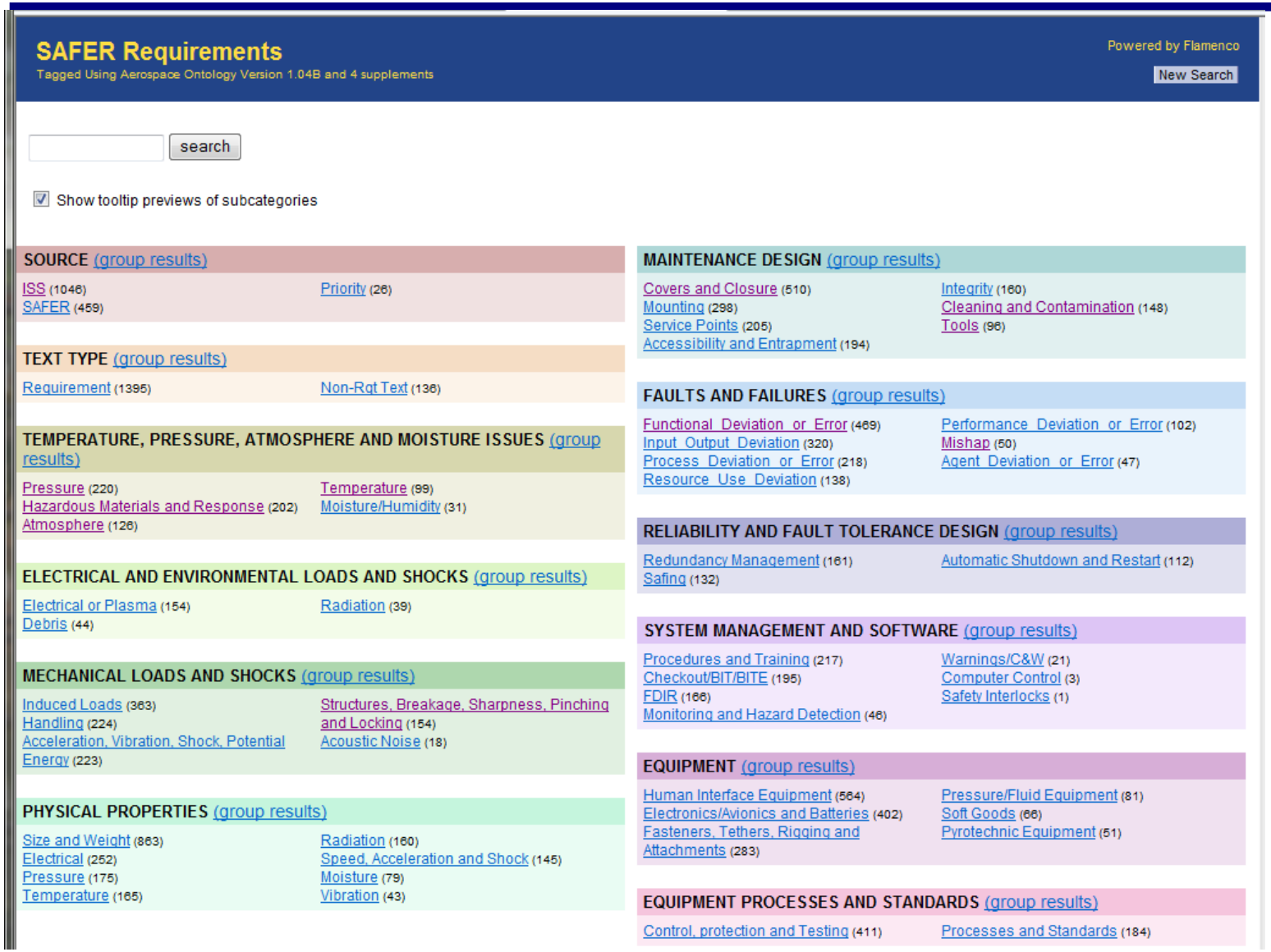
	Avg. Recall	*Avg. Precision*	$\sum T_p$	$\sum F_n$	$\sum F_p$
Search alone	0.54 (0.25-1.00)	0.88 (0.50-1.00)	126	123	25
Text Mining, No tags	0.38 (0.12-0.77)	0.37 (0.12-0.81)	81	118	110
Text Mining, STAT tags	0.66 (0.30-1.00)	0.63 (0.29-0.89)	120	79	71

Means and ranges for the 10 cases: loosely connected, traceability error, unfit, out of limits, bad identifier, debris, electrically disordered, stained, not aligned, and failed start

Conclusion

- STAT average recall was at 86% in the first study
  - Textpresso ontology-based text miner for biological literature achieved 62 % recall of facts about worm genomes
- Quantum Text text miner was disappointing when STAT tags were not added
  - STAT tags improved text mining to 66% average recall
- Text mining appeared not to be worth the trouble
  - The best way to maximize Recall (the proportion of true cases retrieved) would be to combine records retrieved by both search and STAT tagging
  - Flamenco+ faceted browsing and search with semantic annotation supports this approach
  - Flamenco+ dynamically produces trend graphs and tables

Flamenco+ for Requirements Review



New Developments

- Users find it easy to use Flamenco+ to explore and quickly focus on interesting problem groups
  - Filter using both original data fields and tags from AO concept hierarchy
  - Distribute Excel file of small filtered set to other analysts
- Extending to other types of problems reports (institutional, software), requirements and safety analyses
  - Specify selected AO subhierarchies in Flamenco+ to focus on important topics for review in a domain
  - Extract and compare groups of similar requirements to find missing requirements or select a set for a designer
  - Extract model information for safety analysis or verification
    - Attach to system architecture visualization