Reconciler: Matching Terse English Phrases

David R Throop

The Boeing Company
Houston, TX
281 483 5396
david.r.throop@boeing.com

Abstract: *We describe a problem – matching items between a pair of data sources which share no common key, but which both describe objects in terse, English-language phrases. We posit a theory of how such disparities occur, give a procedure for matching, and describe a novel software tool, the Reconciler, which automates this task. Results from three case studies from the International Space Station program are discussed.*

Table of Contents

Frequently, we wish to combine data from two datasets, which draw from the same population of objects. Both datasets have English-language descriptions of the objects but we have no database key common to both datasets. One dataset calls a device 'OVHD LIGHT, the other calls it 'LAMP, Ceiling'. A person can read the description and realize the descriptions match the same object; computers generally cannot. When both datasets have hundreds of entries, combining the data from both sources is an arduous, error-prone manual *Reconciliation Process*. The task becomes especially daunting when both datasets are periodically updated.

This paper presents a novel software tool, the Reconciler, which automates this task.

This problem, in various guises, shows up repeatedly in the International Space Station Project. This tool was developed to match ISS data; most of the examples in this paper draw from the domain of reconciling ISS equipment lists.

This task is similar to data-mining natural language texts and to the indexing done by Internet search engines. However, equipment list datasets have distinguishing features. They have no grammatical structure in the usual sense. They contain large numbers of very terse and nonstandard abbreviations. Many entries differ only by a few characters. Many of their words have attached single characters serving as identifiers.

Causes of Data Divergence – Reconciliation would be unnecessary if everyone used common sources of parts nomenclature and IDs. But there are compelling reasons this doesn't happen. Teams must reference parts before an official name has been generated. Official names change over time and cannot be immediately updated in all applications.

Table 1 - Causes of Variation in Entries

Transformation		Example
Punctuation, spacing & case		"MLI, Berthing Mechanism Blankets " *v* MLI BERTHING MECHANISM BLANKETS
Commentary		RPC 16 (SPARE) *v* RPC 16; MDM *v* MDM – CRITICAL ITEM HANDRAIL *v* HANDRAIL; RELOCATED AFTER 13A
Missing, dropped info		
	Location, system, function, etc, info	US LAB ECLSS ARS RACK 3 CDRA SORBENT BED *v* SORBENT BED
	Modifiers	TETHER *v* ADJUSTABLE EQUIPMENT TETHER
	'Grouping' words	LOAD TRANSFER *v* LOAD TRANSFER UNIT; DRIVE LOCK *v* DRIVE LOCK ASSEMBLY
	Type / Instanced differences	HOSE (2) *v* SUPPLY HOSE AND RETURN HOSE
Misspellings		
	Dropped or added letters, substitutions, transpositions	VLAVE *v* VALVE; OCCASSIONAL *v* OCCASIONAL; TRANSPANDER *v* TRANSPONER; HIEGHT *v* HEIGHT

	Common misspellings, seen repeatedly	GUAGE *v* GAUGE;
		CONTROLLER *v* CONTROLER;

	Variant spellings	GRAY *v* GREY; MODELLING *v* MODELING
	Phonetic spelling	LIGHT *v* LITE; THRU *v* THROUGH
Abbreviations		
	Truncation	STRUCT *v* STRUCTURE
	Omitted Letters	VLV *v* VALVE; FML *v* FEMALE
	X prefix	XPONDER *v* TRANSPONDER; XTRA *v* EXTRA
Root Forms		
	Stems of plurals, gerunds, numbered and tensed verbs	ASSEMBLIES *v* ASSEMBLY; LOCKING *v* LOCK; CONTROLED *v* CONTROL; TAKES *v* TOOK *v* TAKE
	Prefixed and suffixed	DISCONNECT *v* UNCONNECTED; MINIWRENCH *v* WRENCH
	Verb roots of noun forms	ACTIVATE *v* ACTIVATION; RECEIVE *v* RECEIVER *v* RECEPTION *v* RECEPTACLE
Acronyms		
	From official acronym lists	ADCU AC to DC converter unit AIDS airborne integration data system
	Unofficial acronyms 'found' in text	ABC Audio Bus Coupler
	Alternative acronyms for same item	ECLSS (Environmental Control and Life Support Systems) *v* ECLS (Environmental Control and Life Support)
	Redundant acronyms	HIGH RATE MODEM (HRM)
	With punctuation, special chars	GN&C *v* GUIDANCE, NAVIGATION AND CONTROL; C.D.T. *v* CENTRAL DAYLIGHT TIME; P/L *v* PAYLOAD; °F *v* DEGREE FARENHEIT

Transformation	Example
Agglutinations	
Compound words	BACKFLOW *v* BACK FLOW
With abbreviation	FLWMTR *v* FLOW METER
Incorrect split	CAPTUREL ATCH *v* CAPTURE LATCH; V ALVE
Combined with modifier	PUMP3 *v* PUMP #3
Alternate nomenclature	LAF3 (Engineering) *v* LAD3 (Operations)
Alpha / Numeric	
Alpha forms of numbers	THREE *v* 3 *v* III; QUARTER *v* 0.25
Ordinal / Cardinal	SECOND CYCLE *v* 2ND CYCLE *v* CYCLE 2
Descriptives of numbers	DUAL STAGE PUMP *v* TWO STAGE PUMP
'Cute' numerics	S2SR *v* SPACE-TO-SPACE RADIO
0 / O, 1 / l substitution	H2O *v* H20; ATU-l *v* ATU-1
Chemical names	H2O *v* WATER; NH3 *v* AMMONIA
Technical / Common names	TRD *v* TEMPERATURE SENSOR
Synonyms	LIGHT *v* LAMP *v* LUMINAIRE; BACTERIAL *v* MICROBIAL; STOWAGE *v* STORAGE
Hierarchical differences	ARS POWER *v* CDRA POWER *(ARS is the parent of CDRA)*
Specificity differences	ESSMDM *v* MDM *(MDMs are computers; ESSMDMs are enhanced MDMs.)*
Aggregation, ranges	SWITCHES 10 thru 14 *v* SWITCH 12
Dropping leading 0's	LOCATION A052 *v* LOCATION A52
Word Order	MDM C&C-2 *v* C&C-2 MDM
Part Number	
Succession	115271-513 HOSE ASSEMBLY *v* 115271-522 HOSE ASSEMBLY
Different suppliers	1F93224-1 RADIATOR ASSY- ORU, MDM *v* 83-39400-101 RADIATOR ASSY- ORU, MDM
Differences in translation to English	SEG32107059-301 CTB, SINGLE, RUSSIAN FOOD SEG32107059-301 RUSSIAN FOOD CONTAINERS
Completely wrong - copy and edit errors, changes not propagated, data entry problems, database failure	683-00999-2 HALFINCH FCV HOUSING *v* 683-00999-2 FLUID DEGREASER

Most items exist in multiple hierarchies (functional, physical location, power distribution, data distribution…) and users prefer names that provide information about with their particular tasks. Some tasks must call all identical parts by the same name; others require a unique name for each instance. Users drop identifiers from names when its information is obvious from context, and add commentary to names when it's helpful to their tasks. Many COTS tools have no way of importing nomenclature from outside databases and no extra 'slot' for storing program IDs. Many software tools still impose short character-length limits on descriptive fields.

Even when tools don't impose such limits, practical considerations (such as fitting text into crowded diagrams, spreadsheets or small computer displays) drive users to abbreviate. The ISS program integrates components from what were originally three different major contractors, a host of subcontractors, the US government, and foreign suppliers (arriving with part-numbers and descriptions in Cyrillic.) It has faced more such challenges than usual.

Names change as data passes from hand to hand. Table 1 shows a long but tractable list of typical ways that names get transformed. The Reconciliation process recognizes the results of these transforms and accounts for them. We discuss these transformations, roughly in order of their frequency.

Differences *in punctuation, capitalization and spacing* are the largest sources of variation in the ISS datasets. These differences are mostly ignorable, but we use caution matching a .5" PIPE to a 5 INCH PIPE. Commentary is next most frequent – words *about* the item, rather than its name. Following that are differences in *location, system and function* information. Often this information will be obvious or implicit in one data set. Its inclusion leads to long names, so it's omitted. Simple *spelling errors* are the fourth richest source of variation. A library can capture some common misspellings, along with variant and common phonetic spellings, but others must be noted on the fly. Substitution of O for 0 and l for 1 are similar to misspelling but common and pernicious enough to warrant a separate category.

When long phrases get shortened to *acronyms*, program glossaries and acronym lists can help, but are not sufficient. In simple cases, the acronym is the first letter of each word. But articles, propositions and conjunctions are optionally left out of acronyms. Compound words may contribute more letters; PAYLOAD BAY becomes PLB, but PAYLOAD INTERFACE ADAPTOR becomes PIA. The handling of numbers is especially irregular – 'TTS2' is the "temporary threshold shift measured 2 minutes after exposure," [6]. Many acronyms include significant punctuation and some include special characters. Acronyms are *redundant* when they are included in the entry along with the spelled-out form.

Instead of having two names for the same item, some naming variance stems from *hierarchical difference*. Consider a single sensor on an otherwise unpowered device. The power load may be listed as either the sensor or the device – a choice of two different levels in the PART-OF hierarchy. *Specificity differences* reflect different choices in an A-KIND-OF hierarchy. Items exist in multiple hierarchies. A given hose is a kind-of jumper (which is a kind-of connector), structurally part-of the Cabin and functionally part-of the cooling system. Similarly, one-to-many connections cause *aggregation* mismatches, where a single entry (SWITCH 12) must be matched to a range (SWITCHES 10 THRU 14.)

Different work groups evolve their own nomenclatures. On ISS, what Engineering calls *Floor,* Operations call *Deck.* The location codes differ; in the US Lab, Engineering's LAF3 is Operations' LAD3.

Reconciliation is much easier when data includes *Part Numbers.* Part numbers are standardized and eliminate much of the guesswork in matching data from different sources. But part numbers designate a type of part – they don't distinguish different instances of the same part. They also change in particular ways. Each part number has a prefix that identifies the manufacturer or supplier, a second field unique to the part, and a dash number showing the version of the part. The dash number changes when a part's design changes, but not the other fields. Datasets generated at different times should match on the part number base, but may show different dash numbers. There are other sources of part number mismatch. The original manufacturer of a common item gives it their part number. It is integrated into several different assemblies built by other companies and the government; each assigns it their own new part number (which will not necessarily change if they switch suppliers.) When companies are acquired, they change their part numbering systems. Foreign part numbers in Cyrillic or other non-Latin alphabets cause their own set of problems.

Reconciling Datasets – After two datasets have diverged, they frequently there is need to recombine them. The automated reconciliation process works in six steps: Reading and Cleaning, Parsing, Candidating, Matching, Scoring and Reporting.

Reconciler *Reads and Cleans* the entries. In a typical reconciliation problem, the *query* list is shorter or more specific, and we seek the best match among the entries of the other *library* list. The query and library sources are typically tab-delimited files from Excel spreadsheets, (but we've extracted from MS-Word, Acrobat and PostScript.) Data from multiple spreadsheet columns combine into a single entry. The inputs are cleaned of special characters, uppercased and split into tokens.

All the query and library entries are then *Parsed*, breaking them into *tokens* (words, letters, short phrases.) Recognized abbreviations, acronyms, misspellings and stemmed tokens drive to their canonical form. Adjacent tokens combine into single tokens, guided by the information in the *Types* and *Numbered Items* knowledge bases. Tokens are classified by *Group.* An index for each dataset holds the location of the 'interesting' tokens.

For each query, a list of library *Candidates* that share at least one 'interesting' token is generated and then pared, because performing a detailed match between each query and all the library entries would be inefficient. Many datasets restrict the candidates further using domain knowledge.

Matching is performed between the query and each candidate, involving several rounds of token-by-token comparison. Each round applies a *matching filter.* The first filter seeks pairs of tokens that match exactly. Later filters have tests for each of the variations listed in Table 1. The *Standard Groupmatch Score* knowledge base awards a numeric partial-score for each matched token-pair. Matched equipment names score high, matched prepositions score low. Token-matches can be penalized based on the variation – the Location Code 'A57' matches 'A057' but at a lesser score than a match to another 'A57'. Later filters match multiple tokens from one side to a single token on the other.

The matching filters can be adjusted and supplemented for different data sets. For example, the query and library datasets have no common key. But data normally comes with its own key which often encodes useful information. In such cases a filter is added at the end of the matching that decodes the key and matches it to unmatched tokens on the other side.

Scoring sums these partial-scores, counts a penalty for unmatched tokens (again, based on the token's Group) and counts bonuses or penalties for properties of the entire match (e.g., differences in word order.) A pair is disqualified if there are pairs of unmatched tokens from *Mutually Exclusive Groups* – TCS AMMONIA RETURN can't match TCS AMMONIA SUPPLY. The matched candidates are sorted by score, presenting the top-scoring candidate as the best match and noting any ties. If no candidate has a positive score, the query remains unmatched.

Table 2 Reconciler Knowledge bases

Knowledge Base	Description	Example
Acronyms, official	Program approved list of acronyms and abbreviations	CTV← Crew Transfer Vehicle
Acronyms	Other acronyms found in datasets	EEL ← Emergency Egress Light
Ambiguities	Abbreviations with multiple meanings	N2 (Node 2, Nitrogen) COMM (Command, Common, Communication, Commode)

Canonical Forms	Preferred form for variant spellings, variant terminology, common abbreviations of single words	ADAPTER ← ADAPTOR
		FOREWARD ← FWD
		LAB ← USL, USLAB
		TRANSCEIVER ←XCVR

Groups	Parts of speech, semantic tag	NUMBER ← Hundred
		EQUIPMENT ← Pump
		TOOL ← Wrench
		CONNECTION ← Shunt
		PREPOSITION ← Among

Hierarchy	Functional, physical PART-OF relationships; KIND-OF relationships. Supplemented for particular datasets.	TCS ∈ Internal TCS, External TCS ECLSS ∈ ARS, FDS, TCCS, MCA, VACUUM…
Misspellings	Commonly misspelled words	Gauge ← Guage
Mutually Exclusive Groups	If each entry has a different, unmatched member, it's not a match, no matter what.	TOGGLE (Activate, Deactivate) DIRECTION (Aft, Nadir, Port, Forward, Zenith, Starboard)

Nounforms of Verbs	Nouns derived from verbs	ACTIVATE ← Activation
		ARRANGE ← Arrangement
		IDENTIFY ← Identification

Numbered Items	Items that normally have a numeric modifier	ATU (ATU-1, ATU-2) NODE (Node 1, Node 2)
Parts of Speech	Classifies groups	NOUN (Tool, Activity, Connection, Equipment..) ADJECTIVE (Color, Orientation, Timing…)

Standard Group-match Score	Points scored for a match for this group	EQUIPMENT ← 15
		NUMBER ← 3
		ISS_SUBSYSTEM ← 10

Synonyms	Equivalent terms	Light v Lamp ← Luminaire
		Bacterial ← Microbial
		Stowage ← Storage

Types	Specific types of a general class	CABLE (Coaxial, Jumper, …) EXCHANGER (Heat) ENERGY (Radiant, Kinetic, Electrical, Thermal, …)

Knowledge Base	Description	Example

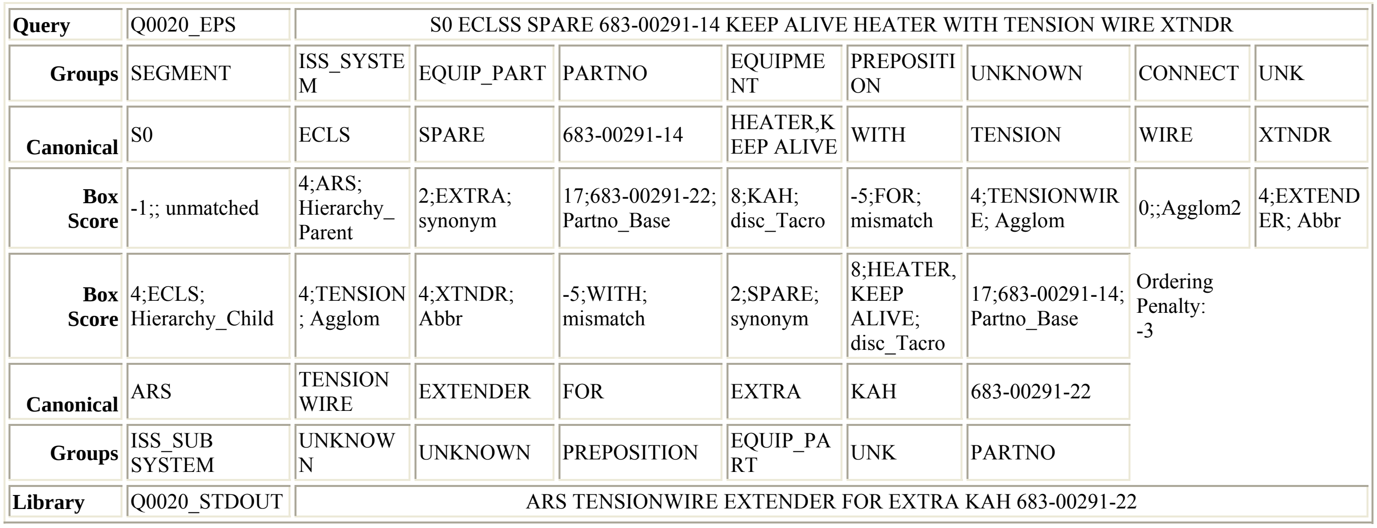
Unmatched Penalties	Penalties scored for an unmatched token of this group. May be adjusted by dataset	EQUIPMENT ← 10
		NUMBER ← 1
		ISS_SUBSYSTEM ← 6

Reconciler then scores the best match qualitatively. A long entry, with many matches and many unmatched tokens, may have a high score but still be a problem match; a short entry can only get a low numeric score. The qualitative scoring depends upon the number of matched, unmatched and mismatched tokens in each group. E.g., Excellent matches must contain at least one matched Equipment, and no match can be Excellent if it contains any unmatched Equipment token. Qualitative scoring is tuned for each dataset.

Reporting presents the matches, and performs and reports analyses that range over the entire set of matches – including library entries that were unmatched or multiply matched, tokens that were not recognized, new acronyms discovered, and stats for ties, average scores, and the qualitative scores.

Application and Results – Table 3 shows actual output for a synthetic match between two entries, demonstrating many of Reconciler's matching features. The entries are '*S0 ECLSS Spare 683-00291-14 Keep Alive Heater with Tension Wire XTNDR*' and '*ARS Tensionwire Extender for Extra KAH 683-00291-2.*' The first line gives the Query entry. The second and third lines show how the Query tokens were grouped and canonized. Coming up from the bottom, the eighth line gives the Library entry; the sixth and seventh show the library groups and canons.

Table 3 Synthetic Example showing matching features



Query	Q0020_EPS	S0 ECLSS SPARE 683-00291-14 KEEP ALIVE HEATER WITH TENSION WIRE XTNDR							
Groups	SEGMENT	ISS_SYSTEM	EQUIP_PART	PARTNO	EQUIPMENT	PREPOSITION	UNKNOWN	CONNECT	UNK
Canonical	S0	ECLS	SPARE	683-00291-14	HEATER,KEEP ALIVE	WITH	TENSION	WIRE	XTNDR
Box Score	-1;; unmatched	4;ARS; Hierarchy_Parent	2;EXTRA; synonym	17;683-00291-22; Partno_Base	8;KAH; disc_Tacro	-5;FOR; mismatch	4;TENSIONWIRE; Agglom	0;;Agglom2	4;EXTENDER; Abbr
Box Score	4;ECLS; Hierarchy_Child	4;TENSION; Agglom	4;XTNDR; Abbr	-5;WITH; mismatch	2;SPARE; synonym	8;HEATER, KEEP ALIVE; disc_Tacro	17;683-00291-14; Partno_Base	Ordering Penalty: -3	
Canonical	ARS	TENSION WIRE	EXTENDER	FOR	EXTRA	KAH	683-00291-22		
Groups	ISS_SUB SYSTEM	UNKNOWN	UNKNOWN	PREPOSITION	EQUIP_PART	UNK	PARTNO		
Library	Q0020_STDOUT	ARS TENSIONWIRE EXTENDER FOR EXTRA KAH 683-00291-22							

The matching itself is shown in the middle two lines, labeled Box Score. The *S0* token was never matched and receives a –1 penalty. ECLS was grouped as a ISS_SUBSYSTEM and matched to ARS (Air Revitalization Subsystem) one of its component systems, for 4 points. *Spare* was matched as a synonym to *Extra* for 2 points. The part number bases were matched even though the dashes were off, for 17 points. The tokens *Keep Alive Heater* were canonized to a single token. Although KAH was not an acronym known to Reconciler, it recognized it as an acronym for *Keep Alive Heater*, for 8 points. The prepositions *for* and *with* mismatched for –5. *Tension* and *Wire* together matched *Tensionwire,* for 4 points. *Xtndr* was an abbreviation match to *Extender* for 4 points. The order of the words differed for –3 points, for a sum total of 30 points.

Case Studies – Reconciler has been applied to seven substantial problem data sets. We'll highlight three of these and discuss the others briefly.

Reliability Block Diagrams: The RBDs capture the functional dependencies of an ISS assembly stage as an acyclic directed graph. These graphs are nearly trees, having a root node and leaf nodes; but interior nodes may have multiple parents. The leaves are equipment; the intermediate nodes are functions. Each node has an ID and a description. In 2001, new standards for encoding the IDs and descriptions were implemented. The existing datasets for Flights 11A and 12A (which partially overlapped) were extensively edited. Besides changing the nomenclature, existing nodes were removed and new ones added. No record of the map from the old nomenclature to the new was made. Existing documents, which referenced the old nomenclature, needed to be updated. For Flight 11A, of the 1460 nodes in the new nomenclature, Reconciler matched 1366 to nodes in the older nomenclature. It found matches for 211 of the 318 additional nodes in Flight 12A.

OP01 Flight and System Data Book titles: These books, [3], [4] in some 31 volumes of around 100 pages each, contain the suggested procedures from Boeing, the ISS integration contractor, to NASA's Missions Operations Directorate. Each Flight Book lists procedures to be used on that flight. Procedures that are used in multiple flights are listed in the Systems Data Books, which the Flight Data Books reference using "Refer to" phrases that called out a Systems Book number and volume, section title and sometimes a section reference number. The System Books' procedures similarly referenced sub-procedures. However, some sections had been renamed, renumbered or moved to other books. Another tool gathered all the references, and all the section numbers and headings and found those which matched exactly. Reconciler found 1018 reference where the book and section number were correct but the name was misspelled, 59 references that could not be reconciled, 117 references that were matched to different books or section numbers, and 69 malformed references.

EPS / STDOUT / MCL Join: Most ISS equipment receives power from one channel of an RPCM (a power supply box.) These three datasets all list the ISS electrical loads with RPCM and channel as a key. Each has a free-text description field, which should all nearly match. Reconciler performed pairwise matches on the three descriptions for each key, separating entries matched 'close enough' from those containing contradictory descriptions.

Measuring Match Accuracy – The Reconciler delivers impressive matching, and several customers have testified that it saved them a great deal of work. Quantifying the match accuracy is difficult, though.

The gold standard for judging the adequacy of a reconciliation run would be comparison to an extensive dataset, hand-matched by a domain expert. None of our customers have been interested in performing such a labor-intensive task. The iterative nature of reconciliation complicates measurement. For each of our datasets, looking at missed matches in the initial reconciliation run has elicited additional matching information (such as synonym or acronym lists) or data sources (such as a source for part numbers to supplement the library listing.) There are also data pairs that the experts can't decide whether they a match or not – complicating scoring.

We are attempting to obtain datasets which have already been hand matched, to use as a standard by which to quantify Reconciler's match accuracy.

Implementation, Related and Future Work – Reconciler is implemented in PERL. It writes its outputs as a set of tab-delimited reports which are read as Excel spreadsheets, and which link to a set of HTML reports. It has no GUI.

To illustrate Reconciler's speed, we reconciled a query set of 1444 entries from the ISS RBDs against 1988 library entries from the ISS Master Component List. Running on a 1 GHz Pentium III with 502 MB of main memory, the run performed 44811 pair matches and took 14 min 08 sec.

Related and Future Work – Portions of Reconciler's functionality is realized in various spelling correcting programs. Additionally, Oracle Text [5] uses a similar approach for matching some abbreviations, synonyms and acronyms from user-supplied lists. Reconciler's matching against hierarchical difference is currently implemented only from user-supplied hierarchies. Wordnet [1] contains an enormous corpus of IS-A and KIND-OF hierarchies. We are investigating calling Wordnet to guide hierarchical matching. We are also in conversation with David Maluf's team about how Reconciler might serve as a matching utility for the Netmark tool [2].

Reconciler will be made more user-friendly. Rather than developing an entire GUI, it may make more sense to integrate Reconciler into a spreadsheet or database program, as most of its input files are developed in those environments.

Conclusions – Reconciling terse lists is a recurring problem that presents recurring challenges. Many features of work processes insure that terminology variation is inevitable; mandating that different work groups use a common terminology will not solve the problem. However, there are a tractable number of general ways in which terminologies depart. It is feasible to dedicate software filters for each of the departures. Each application domain will also have its own departures; writing filters for these is also practical. Therefore, the reconciliation task is automatable. We have demonstrated such automation with the Reconciler and applied it successfully to several domains. We will continue to develop the reconciliation techniques and to integrate them with other list and data tools.

References

[1] Christiane Felibaum, *Wordnet, An Electronic Lexical Database,* MIT Press, 1998, ISBN 0-262-06197-X.

[2] Maluf, David A., Bell, David G., McDermott, Bill, et al, XDB-IPG: *An Extensible Database Architecture for an Information Grid of Heterogeneous and Distributed Information Sources, Information Management*: XDB-IPG-0.9, Seattle Washington, Conference Proceedings, 2003

[3] Operations Data Development & Integration Integrated Product Team*, OP-01 Flight Reports*, http://iss-www.jsc.nasa.gov/ss/issapt/oddi/flt_reports.html

[4] Operations Data Development & Integration Integrated Product Team*, OP-01 System Data Books*, http://iss-www.jsc.nasa.gov/ss/issapt/oddi/sys_book.html

[5] Oracle Technology Network, *Oracle Text – an Oracle Technical White Paper,* March 2002, http://otn.oracle.com/products/text/pdf/10gR1text_twp_f.pdf , Oracle Corp, Redwood Shores CA.

[6] Node Control Software Team, *ACRONYMS/ABBREVIATIONS LIST FOR THE INTERNATIONAL SPACE STATION PROGRAM*, revised February 2002, http://iss-www.jsc.nasa.gov/ss/ issapt/vehipt/sspt7ipt/cdhipt/ccipt/grp/ncsdev/docs/acronyms.doc

Biography

David R. Throop has been an Artificial Intelligence Specialist with The Boeing Company since 1992. He provides engineering software support in the Intelligent Systems Branch in the Automation, Robotics and Simulation Division in the Engineering Directorate at NASA Johnson Space Center. He oversaw development of FMEA modeling software and its use for the International Space Station. His 1979 Bachelors of Chemical Engineering is from Georgia Tech. .His 1992 Ph.D. in Computer Science is from the University of Texas, with a dissertation on Model Based Diagnosis.