Finding the Right Balance Between Automated and Visual QA

Naveh Greenberg,
Director, U.S. Defense Development,
Data Conversion Laboratory

Agenda

- QA&QC, what's the difference?
- Common misconception.
- Examples
- Q&A

QA&QC, what's the difference?

Common misconception

- "My data is all over the place. There's no way to automate anything"
- "My data is very consistent and clean"
- "If it looks good it's tagged correctly"
- "It validates therefor it's good".
- "I don't need to be involved, that's why I hired the experts."
- "I will manually tag it and that will save me the QA time"
- "Just get it to SGML/XML, we'll fix it later"

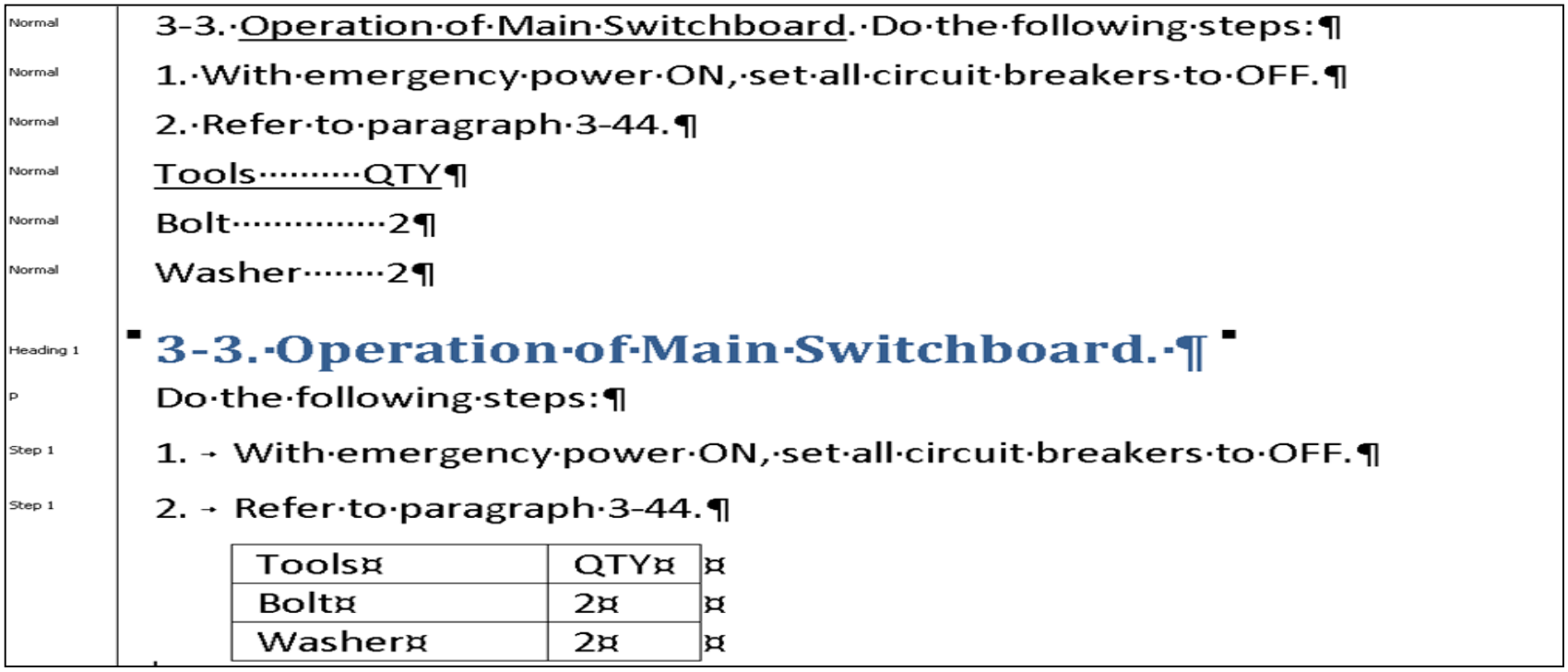
Most Importantly – Plan!!!

Same data but different structure – Analyze the data

Printed Reference:

3-3. Operation of Main Switchboard. Do the following steps:

1. With emergency power ON, set all circuit breakers to OFF.
2. Refer to paragraph 3-44.



Normal 3-3.·Operation·of·Main·Switchboard.·Do·the·following·steps:¶

Normal 1.·With·emergency·power·ON,·set·all·circuit·breakers·to·OFF.¶

Normal 2.·Refer·to·paragraph·3-44.¶

Normal Tools·····QTY¶

Normal Bolt·····2¶

Normal Washer·····2¶

Heading 1 **3-3.·Operation·of·Main·Switchboard.·¶**

p Do·the·following·steps:¶

Step 1 1. → With·emergency·power·ON,·set·all·circuit·breakers·to·OFF.¶

Step 1 2. → Refer·to·paragraph·3-44.¶

Tools¤	QTY¤	¤
Bolt¤	2¤	¤
Washer¤	2¤	¤

Same data but different structure – Analyze the data

Printed Reference:

3-3. Operation of Main Switchboard. Do the following steps:

1. With emergency power ON, set all circuit breakers to OFF.
2. Refer to paragraph 3-44.

Example 1 – This will parse:

<para>3-3.<emphasis type="uline">Operation of Main Switchboard.</emphasis> Do the following steps: </para>
<para>1. With emergency power ON, set all circuit breakers to OFF. </para>
<para>2. Refer to paragraph 3-44.</para>

Example 2 – so will this:

<para>3-3. <emphasis type="uline">
Operation of Main Switchboard.</emphasis>
Do the following steps: <seqlist>
<item>With emergency power ON, set all circuit breakers to OFF. </item>
<item>Refer to paragraph <xref xrefid="par3-44“/>. </item></seqlist></para>

Example 3 – and this:

<para0 id="par3-3">
<title>Operation of Main Switchboard.</title>
<para>Do the following steps:</para>
<step1><para>With emergency power ON, set all circuit breakers to OFF. </para></step1>
<step1><para>Refer to <xref xrefid="par3-44“/>.</para></step1></para0>

Tagging Structure Completeness Check

Even if it parses cleanly, manually checking if data tagged properly according to the tagging rules is very difficult and time consuming so you need to rely on automation:

- Paragraph labels and hierarchy
- Cross-references
- Lists vs. steps

Examples:

This will parse:

```
<para>3-3.<emphasis type="uline">Operation of Main
Switchboard.</emphasis> Do the following steps: </para>
<para>1. With emergency power ON, set all circuit breakers to OFF. </para>
<para>2. Refer to paragraph 3-44.</para>
```

And so will this:

```
<para>3-3. <emphasis type="uline"> Operation of Main Switchboard.</emphasis>
Do the following steps: <seqlist>
<item>With emergency power ON, set all circuit breakers to OFF. </item>
<item>Refer to paragraph <xref xrefid="par3-44”/>. </item></seqlist></para>
```

Text (Extraction required different QA process than migration)

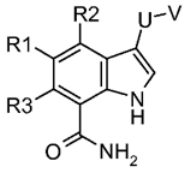
Sample Page

where R1, R2, R3, U and V are defined below and to pharmaceutically acceptable salts thereof.

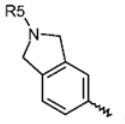
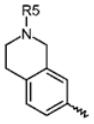
The compounds of the invention are inhibitors of IKK2 and can be useful in the treatment of disorders associated with inappropriate IKK2 (also known as IKKβ) activity, such as rheumatoid arthritis, asthma, and COPD (chronic obstructive pulmonary disease). Accordingly, the invention is further directed to pharmaceutical compositions comprising a compound of the invention. The invention is still further directed to methods of inhibiting IKK2 activity and treatment of disorders associated therewith using a compound of the invention or a pharmaceutical composition comprising a compound of the invention.

DETAILED DESCRIPTION OF THE INVENTION

The invention is directed to compounds according to formula (I):



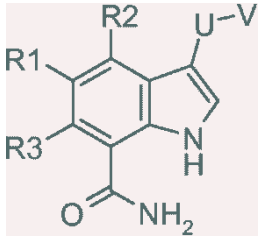
(I)

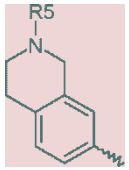
where R1 is the group –XYZ or  or  ;

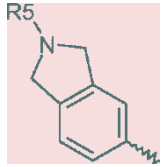
X is phenyl, heteroaryl, 1,2,3,4-tetrahydronaphthalenyl, or 2,3-dihydro-1*H*-indenyl,

Conversion Step 1– Artifact Removal

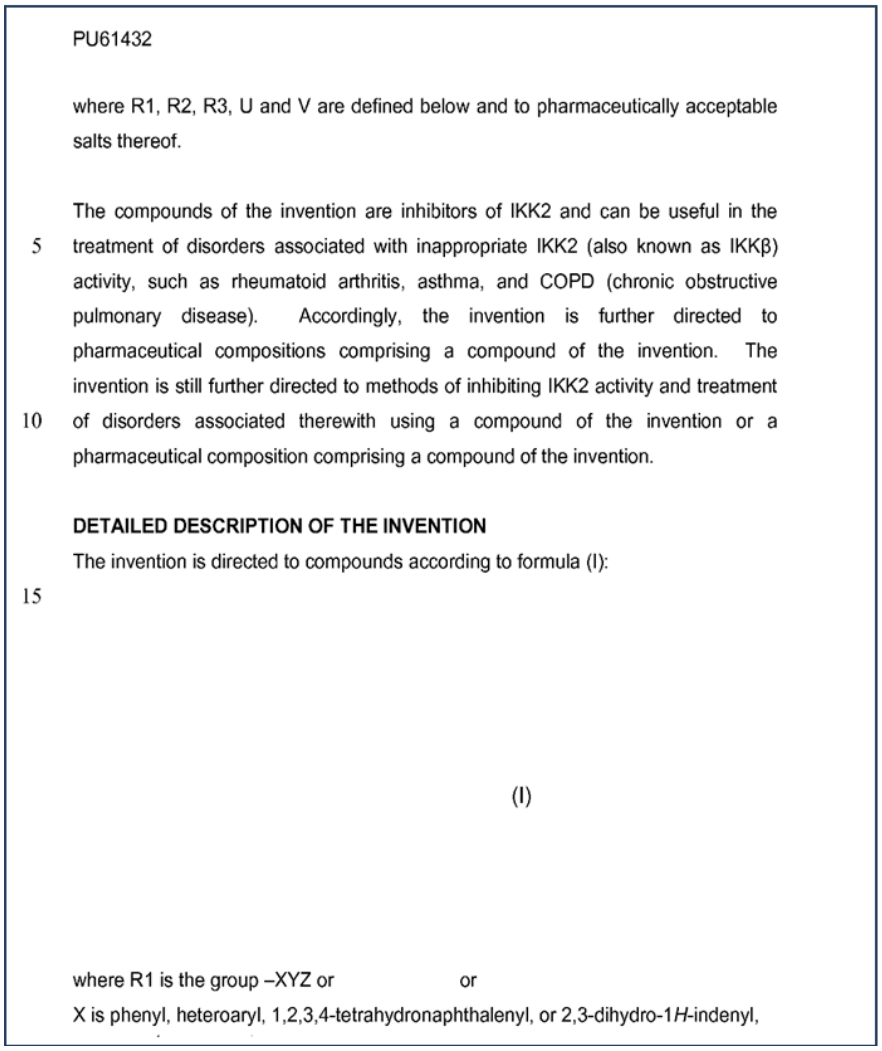
- Extract anything which is not text or can create problems for OCR conversion i.e. extensive math, subscript, superscript.
- It stores the location of that image to help us recreating the document.







Text (Extraction required different QA process than migration)



PU61432

where R1, R2, R3, U and V are defined below and to pharmaceutically acceptable salts thereof.

The compounds of the invention are inhibitors of IKK2 and can be useful in the treatment of disorders associated with inappropriate IKK2 (also known as IKKβ) activity, such as rheumatoid arthritis, asthma, and COPD (chronic obstructive pulmonary disease). Accordingly, the invention is further directed to pharmaceutical compositions comprising a compound of the invention. The invention is still further directed to methods of inhibiting IKK2 activity and treatment of disorders associated therewith using a compound of the invention or a pharmaceutical composition comprising a compound of the invention.

DETAILED DESCRIPTION OF THE INVENTION

The invention is directed to compounds according to formula (I):

(I)

where R1 is the group –XYZ or or

X is phenyl, heteroaryl, 1,2,3,4-tetrahydronaphthalenyl, or 2,3-dihydro-1*H*-indenyl,

Conversion Step 2 - OCR

- Now the OCR solution can process this image accurately (99.6%) for text extraction.

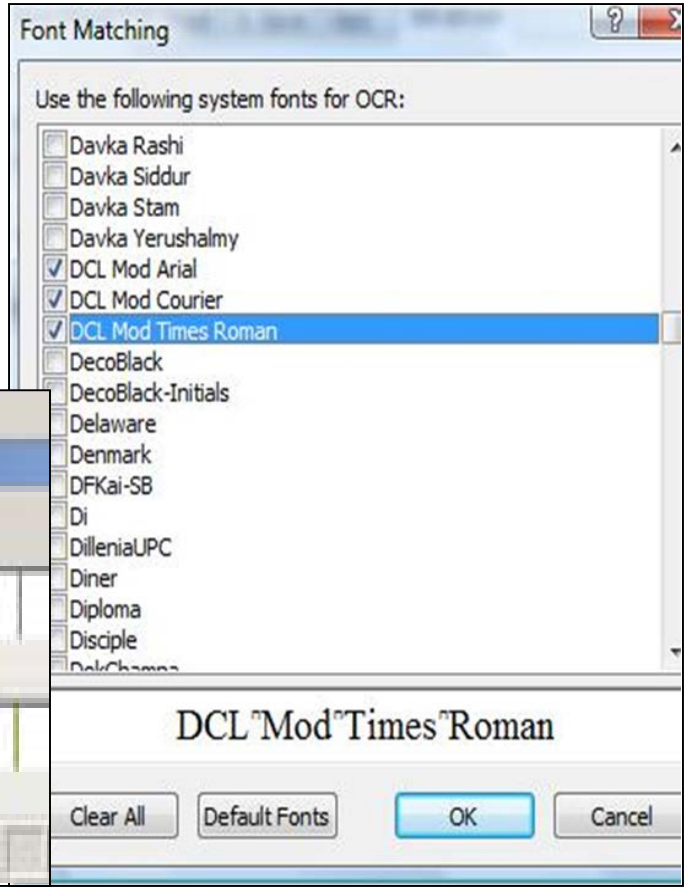
Conversion Step 3 – Post Processing

- Once we have the OCR-extracted text, images and their coordinates, the post processor creates well-formed XML. The XML can be designed to match any XML Schema.

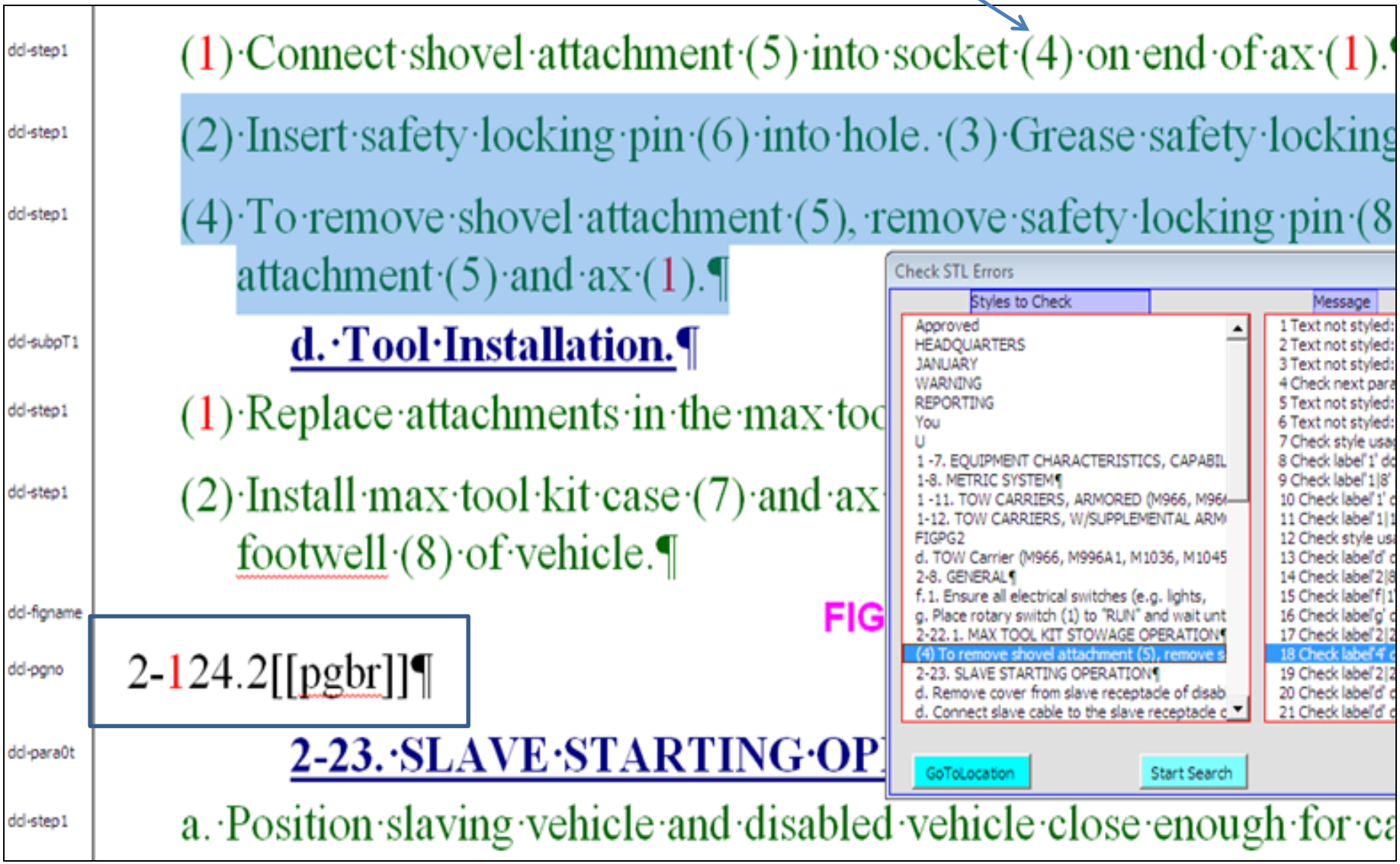
Final Results

- Searchable text in XML format that can be stored and accessed via end-user systems, e.g. front-end, content management.
- Style sheets can regenerate the document in its original form using the output XML and associated images.

Text (Extraction required different QA process than migration)



Automated structure review



Tables – A Manual and Automated process

Table 2. Cable Assembly W443 - Wire List

CFM TJ1 W443 P1	CFM TJ3 W443 P2	MFM TJ2 W443 P3	MFM TJ3 W443 P4	MFM TJ4 W443 P5	SRU J4 W443 P6	SRU J5 W443 P7
P1-1	-	-	-	-	-	P7-13
P1-2	-	-	-	-	-	P7-11
P1-19	-	-	-	-	-	P7-67
P1-23 Jumper	P2-102	-	-	-	P6-91	P7-32
P1-52 Jumper	-	-	-	-	-	-
P1-24	P2-104	-	-	-	P6-90	P7-28
P1-53 Jumper	-	-	-	-	-	-
P1-83 Jumper	-	-	-	-	-	-
P1-25	-	-	-	-	-	P7-46
P1-26	-	-	-	-	-	P7-50
P1-27	-	-	-	-	-	P7-68
P1-29	-	-	-	-	-	-
P1-114 Jumper	-	-	-	-	-	-
P1-30	P2-18	-	-	-	-	P7-36

Table 2. Cable Assembly W443 - Wire List (Continued)

CFM TJ1 W443 P1	CFM TJ3 W443 P2	MFM TJ2 W443 P3	MFM TJ3 W443 P4	MFM TJ4 W443 P5	SRU J4 W443 P6	SRU J5 W443 P7
-	P2-66	-	-	-	-	P7-101
-	P2-78	P3-18	-	-	-	P7-98
-	P2-107 Jumper	P3-24 Jumper	-	-	-	-
-	-	P3-32	-	-	-	-
-	-	P3-36	-	-	-	-
-	P2-79	P3-22	-	-	-	-
-	P2-108 Jumper	P3-30 Jumper	-	-	-	-
-	-	P3-34	-	-	-	-
-	-	P3-38	-	-	-	-
-	P2-83	-	-	-	-	P7-100
-	P2-89	-	-	-	P6-48	P7-76
-	P2-90	-	-	-	P6-19	P7-78
-	-	-	-	-	P6-46 Jumper	-
-	P2-92	-	-	-	-	P7-52
-	P2-98	-	-	-	-	P7-47
-	P2-109	P3-50	-	-	-	-
-	P2-110	P3-10	P4-69	P5-36	-	-
-	-	-	-	P5-38 Jumper	-	-
-	-	-	-	P5-62	-	-
-	P2-111	P3-13	P4-25	-	-	-
-	-	-	P4-29 Jumper	-	-	-
-	P2-124	-	-	-	P6-101	P7-89

Content Reuse Analysis – Automated & visual QA

A17-1_2010_U002_025 line: 2054

engineering test : a test carried out by or witnessed by a registered or licensed professional engineer, testing laboratory, or certifying organization to ensure conformance to Code requirements.

<Exact matches>

A17-1_CSA_B44_Edtn_2013_000505_U100_2013_015 line: 2164

A17-1_Handbook_Edtn_2010_032114_000501_u004_2010_033 line: 19514

A17-3_Edtn_2011_000502_U001_2011_045 line: 7406

90002124_K line: 1731

The XBee® Wi-Fi modules are compatible with other devices that use 802.11 bgn technology. These include Digi external 802.11x devices like the ConnectPort products and the Digi Wi-SP, as well as embedded products like the ConnectCore series and Digi Connect series of products. More information on these Digi products can be found at:

<Exact matches>

90002180_H line: 2100

90002124_K line: 4014

The XBee Wi-Fi RF modules support both UART (Universal Asynchronous Receiver/Transmitter) and SPI slave mode (Serial Peripheral Interface in slave mode only) serial connecti

<Exact matches>

90002180_H line: 4358

Content Reuse Analysis – Automated & visual QA

source file: BE313 line: 122
found 12 paragraphs

warns about hazards that will or can cause minor personal
The label NOTICE indicates special instructions which are important but
Carefully read and follow all safety instructions in this manual
Ground pump before connecting to power supply. Disconnect power before
Wire motor for correct voltage. See "Electrical" section of this
Meet National Electrical Code, Canadian Electrical Code, and local codes
Follow wiring instructions in this manual when connecting motor to
Capacitor voltage may be hazardous. To discharge motor capacitor, hold
Do not touch an operating motor. Modern motors are designed
Do not allow pump or any system component to freeze.
Keep work area clean, uncluttered and properly lighted; store properly
Pump body may explode if used as a booster pump

Matches file: BE745 line: 132
Matches file: S367 line: 156

Content Reuse Analysis – Automated & visual QA

AG_1b_U004_054 line: 14401

FD-5220 OTHER WELDS Fillet welds, butt welds, and seal welds shall be visually inspected in accordance with AA-6331.

FD-5220 {FH-5220} OTHER WELDS Fillet welds, butt welds, and seal welds shall be visually inspected in accordance with AA-6331.

~~FD-5220~~ FH-5220 OTHER WELDS Fillet welds, butt welds, and seal welds shall be ~~visually~~ inspected in accordance with AA-6331.

AG_1b_U004_054 line: 22357

253P6310 line: 90

This is the safety alert symbol. When you see this symbol on your pump or in this manual, look for one of the following signal words and be alert to the potential for personal injury.

This is the safety {safety-alert.} alert symbol. When you see this symbol on your pump or in this manual, look for one of the following signal words and be alert to the potential for personal injury. {injury: injury!}

This is the safety alert symbol. When you see this symbol on your pump or in this manual, look for one of the following signal words and be alert to the potential for personal ~~injury.~~ injury!

S502 line: 69

S506 line: 56

S509 line: 54

This is the ~~safety alert symbol.~~ safety-alert. When you see this symbol on your pump or in this manual, look for one of the following signal words and be alert to the potential for personal ~~injury.~~ injury:

S280_10SP4C02H03 line: 183

S280_10SP4C02J04 line: 183

This is the ~~safety alert symbol.~~ safety-alert. When you see this symbol on your pump or in this manual, look for one of the following signal words and be alert to the potential for personal injury.

PN280 line: 146

S492 line: 105

The Conversion Specification – A Living QA Document

¤	Description¤	Example¤	Inferred·Assumptions¤	Tagging·Examples¤	Open·Items¤
76.	**Cross-references¤**	TO·XXX-XX-XXX· page·37· step·b¤	••Reference to an internal paragraph, figure, foldout, or table. Cross-references to chapters and sections will not be provided.¶ ••Each <xref> will have a "xrefid" attribute assigned as follows: *X*#-#¶ Where *X* is "fig" for a figure; "figfo" for a foldout; "tbl" for a table; "chp" for as chapter; "par" for a paragraph, and #-# is the id of item being referred to.¶ ¶ Notes:¶ - → Cross-references will be identified by text in one of the following formats:¶ Paragraph #, Table #, Figure #, Fig. #, FO-#. The only allowable variations are for text casing and pluralization (e.g., Tables #). Other variations will not be handled.¶ - → The keywords paragraph, table, or figure, and number that follows them, will be replaced by the <xref> element.¶ - → In the case of plural references (e.g. Tables # and #), a cross-reference will be created for each #, the #s can be separated by "," (comma), "thru", or "and".¶ - External cross-references will not be handled.¶ ¶ Cross references to multi-sheet figures will point to the figure.¤	Paragraph #:...refer to paragraph 1-5¶ *becomes:*¶ ...listed in <xref xrefid="par1-5">¶ ¶ Figure #:...listed in Figure 1 or Fig. 1¶ *becomes:*¶ ...listed in <xref xrefid="fig1-1">¶ ¶ FO-#:...(FO-5-1)¶ *becomes:*...(<xref xrefid="figfo5-1">)¶ ¶ Table #:...see Table 1¶ *becomes:*¶ ...see <xref xrefid="tbl2-1">¶ ¶ Ranges:...tables 1 thru 4¶ *becomes:*¶ ...<xref xrefid="tbl2-1"> thru <xref xrefid="tbl2-4">¶ ¶ For multi-sheet figures:...see Figure 6-1¶ becomes: see <xref xrefid="fig6-1a">¤	¤
77.	.....¤	.....¤	••.....¤	.....¤	¤
78.	**Sub·Figure¤**	TO·XXX-XX-XXX· page·2-48¤	••Indicated by a graphic with the caption text: "(Sheet # of #)".¶ ••The first sheet and each subsequent sheet will be treated as <subfig>, with its own <graphic> element.¶ ••The graphic boardno will be the same as the figure id, with a sequential letter appended to each sub-figure.¶ ¶ **Notes:**¶ - → "pgwide" value will default to 1 (should be set to "0" if graphic spans only one of two columns on page)¶ - → "scalefit" value will be set to "1"¶ - → "reprodep" value be set to the actual vertical dimension of the graphic, in inches (in), not to exceed 8.5in¶ - → "reprowid" value will be set to the actual horizontal dimension of the graphic, in inches (in), not to exceed 11in¶ - → Dimensions will be rounded to two decimal places.¤	<subfig id="fig1-1a"><graphic boardno="fig1-1a" reprodep="" reprowid="" scalefit="1"><legend>…</legend></subfig>¶ ¶ <subfig id="fig1-1b"><graphic boardno="fig1-1b" reprodep="" reprowid="" scalefit="1"></subfig>¶ ¤	¤

Content Tagging Validity Checks – Automated process

jg_parts.txt* | 1A-10C-4-2 20070401 wC26.sgm

partno	inIPB	count	phrase
160D116015	---	2	PN 160D116015, for gouges and loose or elongated bushings. Refer
160D116031-5	---	4	On aircraft assembly PN 160D116031-5, do not perform step
160D117100-31	---	2	change="delete">oil</change> to shank of bolt (21). Install spacer PN 160D117100-31
160D117100-31	---	2	PN 160D117100-33 (25) and PN 160D117100-31 (26)

Search Results

Searching for: 160D117100-31

1A-10C-2-51JG-1 20070401 wC24.sgm(4847): PN 160D117100-33 (25) and PN 160D117100-31 (26).</para>

1A-10C-2-51JG-1 20070401 wC24.sgm(4973): change="delete">oil</change> to shank of bolt (21). Install spacer PN

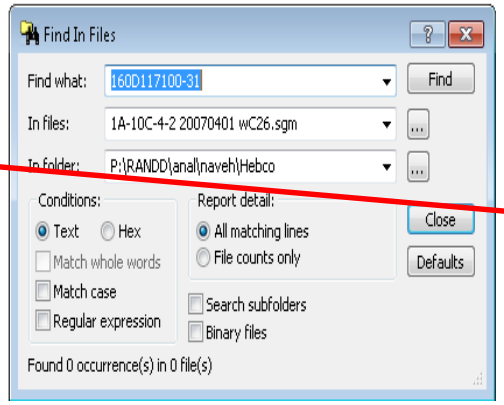
1A-10C-2-51JG-1 20070401 wC25.sgm(4867): PN 160D117100-33 (25) and PN 160D117100-31 (26).</para>

1A-10C-2-51JG-1 20070401 wC25.sgm(4993): change="delete">oil</change> to shank of bolt (21). Install spacer PN

Found 4 occurrence(s) in 2 file(s)

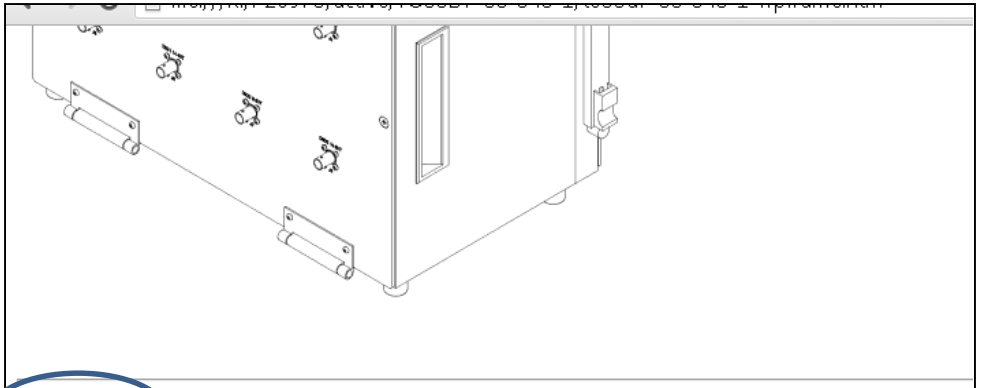
Searching for: 160D117100-31

Found 0 occurrence(s) in 0 file(s)



Part number 160D117100-31 is found in the Job Guides but not in the IPB

Viewing Your Converted Data



[para0] 1-11 **AUTOMATIC CALIBRATION**

Automatic testing of the ARTS is accomplished using the TPIH. Under program control, the TPIH provides ARTS g
PATEC and ARTS self-generated analog and digital stimuli to every ARTS station instrument and monitors response

[para0] 1-13 **RELATED PUBLICATIONS**

Technical orders required to supplement the information within this manual are listed in table 1-5.

[para0] 1-15 **TABLE OF TPS COMPONENTS**

table 1-3 lists the nomenclature, reference designation and description of the TPS. In addition, the Computer Prog
testing. The TPIHis used to verify calibration of the ARTS to NIST standards, to adjust as necessary, and identify an
2 loop-back plugs and one attenuator. The TPS performs functional performance tests, calibration tests, performs ca

[para0] 1-18 **CAPABILITIES AND LIMITATIONS**

TABLE 1-5

Publication List

TECHNICAL ORDER	TECHNICAL ORDER TITLE
TO 00-25-234	General Shop Practice Requirements for the Repair, Maintenance, and Test of Electrical Equipment

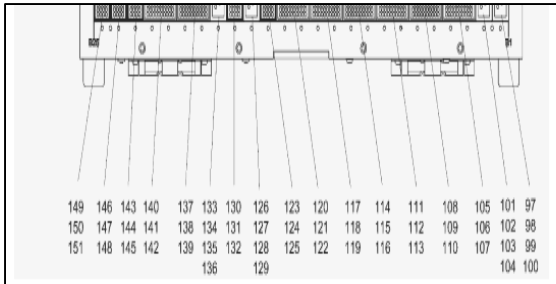


Fig. Index / Sht. #	Part #	CAGE (FSCM)	Description 1 2 3 4 5 6 7	Units per Assy	Usable on Code	SMR Code
7-2	201323711-10	98752	ITA ASSEMBLY, MAINTENANCE	REF		
14	201323733-01	98752	• CONNECTOR, BNC, F-F	8		
15	201323363-01	98752	•• SCREW, MACHINE PAN HEAD	32		
16	NAS620C3L		•• WASHER, FLAT NO 3	32		
17	201323560-01	98752	• LABEL	1		
18	201323709-01	98752	• RECEIVER, ICON960	1		
19	MS16995-10		•• SCREW, CAP, SOCKET HEAD	4		
20	MS35338-135		• WASHER, LOCK-SPRING, HELICAL	4		
21	MS35649-244		•• NUT, PLAIN, HEX	4		
22	201323721-01	98752	• COVER, PROTECTIVE, ITA	1		
7-2-23	MS18029-1S-8		• COVER, ASSEMBLY, ELECTRICAL	2		
24	MS27212-1-8		• TERMINAL BOARD ASSEMBLY (TB1 - TB2)	2		

Viewing Your Converted Data

NOTE

It may be necessary to loosen setscrews 3 and reposition thumbwheel 2 to obtain reading.

END NOTE

Loosen thumbwheel setscrews 3. Move top of thumbwheel 2 towardHOD 4, until mechanical stop is reached.

d. Tighten either setscrew 3. Only one is required.
e. Record ohm meter reading.
f. Move ohm meter lead from contact 2 to contact 3 of connector J3 1.
g. Move top of thumbwheel 2 away fromHOD 4 to the other mechanical stop. Record ohm meter reading.
h. Subtractstep e ohm meter reading fromstep g ohm meter reading.
i. Divide the difference by 2. Add the result (positive or negative) to the ohm reading recorded instep e.
j. Move ohm meter lead from contact 3 to contact 2 of connector J3 1.
k. Move top of thumbwheel 2 towardHOD 4 until resultant reading fromstep i is obtained ±10 ohms.
l. Loosen setscrew 3 on thumbwheel 2. Move top of thumbwheel 2 toward theHOD 4 to the mechanical stop.
m. Tighten both setscrews 3.
n. Record ohm meter reading.
o. Move ohm meter lead from contact 2 to contact 3 of connector J3 1. Move top of thumbwheel 2 away fromHOD 4 until the mechanical stop is reached.
p. If the ohm meter reading is not within 50 ohms of the reading obtained in step n, repeat steps c throughо.

FIGPG4-246

Automated QC

```
****************************
Invocation: airqc($Revision$1.31) -mv:\..
* ERR * Image 'f015.tif' was not scanned
* ERR * Image 'fig4a-2.tif' was not scanned
* ERR * Not standard column '5' width: ''; line #=404; line text=0
* ERR * Not standard column '6' width: ''; line #=404; line text=0
* ERR * graphics\needs no work\pn13a2550-18.tif Image Compression is not Group IV: (Compression ratio: 0.38:1)in   Raw size: 1218  Actual s
* ERR * graphics\needs no work\pn13a2550-01.tif Image not Black and White:   Number of Colours: 256
* WARN * Number of cells '3' more than described in heading '2'in table #1; line #=105; line text=<row>  , # 1   , files: dmc-XXXX-a-e00-00
* WARN * Number of cells '3' more than described in heading '2'in table #1; line #=113; line text=<row>  , # 1   , files: dmc-XXXX-a-e00-00
* WARN * Unexpected content '7-217222500-61 1' in <partno>7-217222500-61 1</partno>; line #=20026, msg=597
* WARN * <title> starts from space ' GENERAL'; line #=66
* WARN * Check if not reference 'Section IV ' in <para>Section IV contains detailed information on the functions of the Circuit Card Test m
* WARN * Check if not ref  ' Steps 3 ' in <para>Repeat Steps 3 and 4 this time adjusting R43 and R20 for 1000.00 &plusmn;0.3 and 10000.0 &p
* WARN * Element has no content '' in <title></title>; line #=1700, msg=606
* WARN * Suspicious bullet at start of para/item '&bull;' in <para>&bull; All values in &lt;xref xrefid="WP016-00-tbl1" displayvalue="Table


****************************************************************************************************************************************************
****************************************************************************************************************************************************


****************************
Invocation: asmefx1($Revision$) -mk:...
* ERR * #e819, Label not rec. at <para1><para>a)<tab>The maximum allowable stresses for SA-182/SA-182M Class F91 in Section II, Part D, Subp
* ERR * #e819, Label not rec. at <para1><para>b)<tab>This Case number shall be shown on the Manufacturer&#x02019;s Data Report.</para></par
completion: # recs: 46, # errs: 2, # warn: 0, # info: 0, Mon Mar  2 16:55:23 2015

****************************
Invocation: tblnotes($Revision$) 2815.f30 -o2815.f40 Mon Mar  2 16:55:23 2015
* WARN * #w303, Did not find 'table' '5A', rlvl='1', referenced on line: ; line #=28; line text=The maximum allowable stresses for SA-182/S

completion: # recs: 0, # errs: 0, # warn: 1, # info: 0, Mon Mar  2 16:55:23 2015

****************************
Invocation: K:\p26488\sw1\asmerefs.exe($Revision: 1.0.8$) -q -fk:\P26488\sw1\tbls\asmerefs.tbl 2815.f60 -o2815.f65  Mon Mar 02 16:55:23 201
* ERR * Repeated tif image 'fig7-25.tif' in declaration
* ERR * Input file 'K:\P26928\Source\conv\images\fig1-1.tif' is not readable
* ERR * dquot not closed at the end of: This column contains the five-digit Joint Military Services Uniform SMR Codes or the Air Force Pecu
* ERR * Width=9.64 not fixed to 7.1 in table: tabmat; line #=0; line text=
* WARN * #w206, <subpara> Label was not recognized in ; line #=901; line text='The disk bay assembly is located in the upper lef'
```

Q&A

Naveh Greenberg
Director, U.S. Defense Development,
Data Conversion Laboratory
(718) 307-5758
ngreenberg@dclab.com
@dclaboratory