Data Format Description Language &

APACHE

Daffodil™

(incubating)

Killing the Data Format Problem Forever

Mike Beckerle, Data Archeologist, Tresys Technology
mbeckerle at tresys.com or at apache.org

TRESYS Deep.

Got EDIFACT Data?

```
UNA:+.?*'
UNB+UNOC:4+5790000274017:14+5708601000836:14+990420:1137+17++INVOIC++++1'
UNH+30+INVOIC:D:03B:UN'
BGM+380+539602'
DTM+137:19990420:102'
RFF+CO:01671727'
NAD+BY+5708601000836::9'
RFF+VA:UK37499919'
NAD+SU++IBM UK'
RFF+VA:UK19430839'
RFF+ADE:00000767'
NAD+DP+++MyCompany+MyStreet+MyTown++1234+UK'
CUX+2:GBP:9'
LIN+1++V0370246:IN'
IMD+F++:::Collectors edition of The Hobbit with Tolkien?'s original colours on sleeve'
QTY+47:5:PCE'
MOA+66:49.15:GBP'
PRI+AAA:9.83:CT::1:PCE'
RFF+CO:01671727:1'
ALC+C'
MOA+23:13.6:GBP'
LIN+2++:IN'
```

Got bit-packed binary data ?

- Bytes are

  `09 20 42 F0 0D B8 DD`

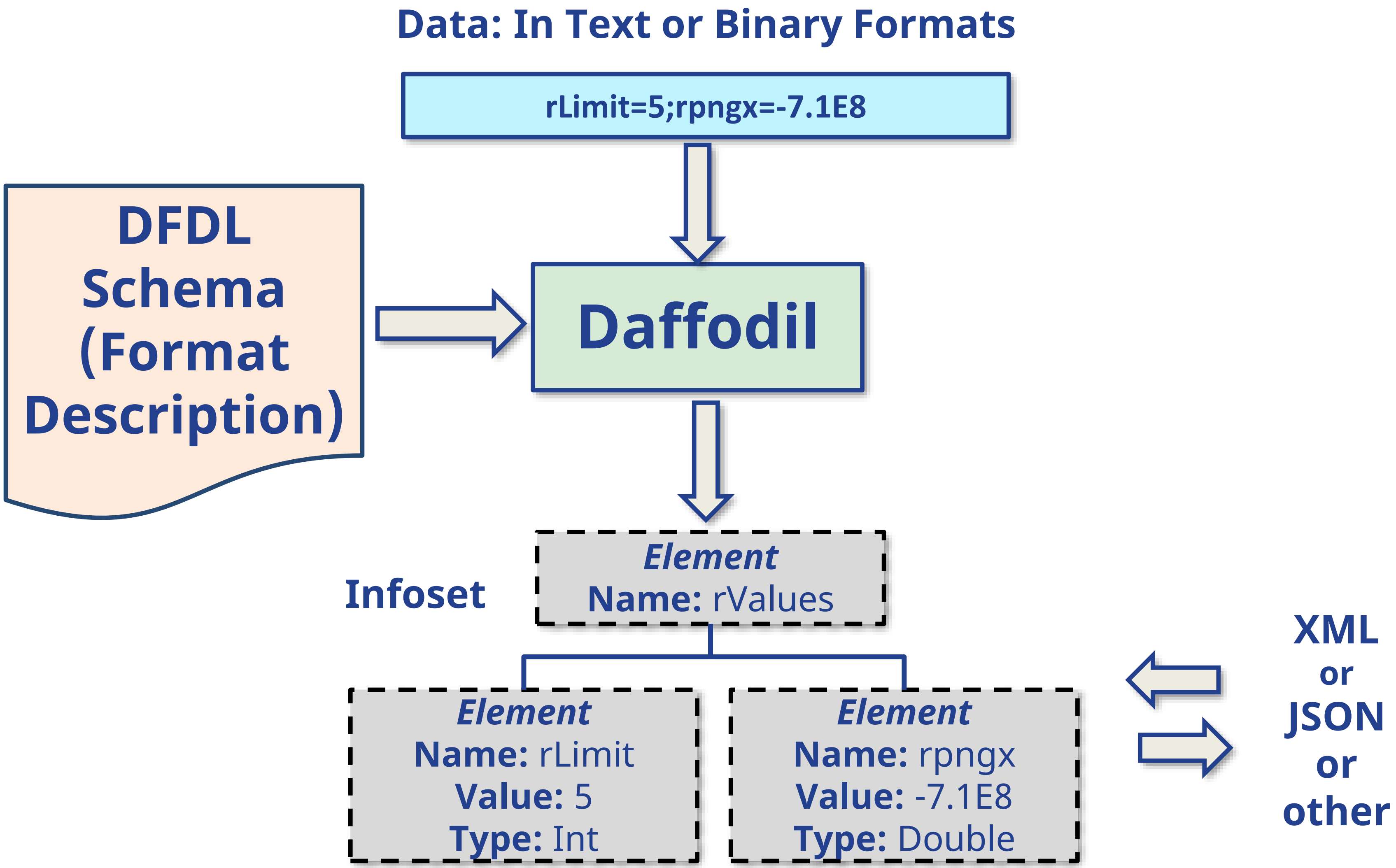
- Fields are decribed as:

Message Number	XXXXXX00	00001xxx
FPI for Message Subtype		XXXXX0xx
FPI for File Name		XXXX0xxx
FPI for Message Size		XXX0XXXX
Operation Indicator		X01XXXXX
Retransmit Indictor		0XXXXXXX
Message Precendence Codes		XXXXX010
Security Classification		XXX00XXX
FPI for C/R marking		XX0XXXXX
GPI for Orig DTG		X1XXXXXX
Year	XX110000	0XXXXXXX
Month	XXXXXX01	11XXXXXX
Day		X00011XX
Hour	XXXX1000	0XXXXXXX

Got NACHA Data?

```
101 121000248 1210002480608080107A094101WFB-W EDI CUST. DATA   WFB-E ACH SPINAT DATA
5200APD TX/FINCL SVC323413684           9666666606CCDAPD - TAX 0608020608022141021000024030649
62710700543200004001191477   0542151200614007046488KD8BRODY LABORATORIES INCFS0021000024030840
820000000100107005430005421512000000000000009666666606                         021000024030649
5200ARAR                                9000290001CCD PAYMENT        0608072191021000027294149
62205100141200004945037059   00004036762394128        VIA LICENSING CORP      0021000027294283
820000000100051001410000000000000000004036769000290001                         021000027294149
5200ACME   WORLDWIDEDIRECT DEPOSIT      9954245682CCDA/P             0608022141122000030000219
62210700543200004001191477   000005000010000142611008 100815047371712006      0122000036548030
62210700543200004001191477   000014750010000142611008 100815047375772006      0122000036548031
820000000200214010860000000000000000001975009954245682                         122000030000219
5200BEST BANK NA    5046001042958       9560900031CCDEDI PAYMNT      0511181231021101100000014
6220510014149995275638771    0000037500504058967      XXXXXXXX BRANCH INC     1021101100001681
705RMR*OI*0140611**-1170.49*25*1170.36*                                            00010001681
6220610002279990124617283    0023519545504058968      XXXXXXXXXXXXX SERVICE   1021101100001682
705RMR*ADG*504058968                                                               00010001682
6220211011089997375415088    0000690240502397811      XXXXXX INC              0021101100001683
6220510014149998010236916    0000662215502397782      XXXXXXXXXXXXXXXXXXXXXX  0021101100001684
820000000600184104140000000000000000249095009560900031                         021101100000014
5200SCISSORS                            4042896127CCDBATCH     0808060608082201011200360001437
6220910000190001065201       0100813537101017         5M CORPORATION          0011200362667919
820000000100091000010000000000000001008135374042896127                         011200360001437
9000005000003000000110064712185000542151200000126324213
9999999999999999999999999999999999999999999999999999999999999999999999999999999999999999999999
```

Parse Data based on DFDL Schema



Data: In Text or Binary Formats

rLimit=5;rpngx=-7.1E8

DFDL Schema (Format Description)

Daffodil

Infoset

Element
Name: rValues

Element
Name: rLimit
Value: 5
Type: Int

Element
Name: rpngx
Value: -7.1E8
Type: Double

XML or JSON or other

Data Format Description Language

DFDL →DaFfoDiL

- DFDL is a way of describing data formats
- It is NOT a data format itself!

- Open Standard from the Open Grid Forum (OGF)
- DFDL Specification - Version 1.0 – Sept. 2014
- 2 other DFDL Implementations (IBM, ESA)

- DFDL has union of capabilities across many marketplace data integration products/tools/packages

Convert that NACHA to XML Please....

```xml
<ACHFile xmlns="ach:2013">
<FileHeaderRecord>
<RecordTypeCode>1</RecordTypeCode>
 <PriorityCode>01</PriorityCode>
<ImmediateDestination> 123456789</ImmediateDestination>
<ImmediateOrigin> 987654321</ImmediateOrigin>
<FileCreationDate>071030</FileCreationDate>
<FileCreationTime>1634</FileCreationTime>
<FileIdModifier>A</FileIdModifier>
 <RecordSize>094</RecordSize>
<BlockingFactor>10</BlockingFactor>
 <FormatCode>1</FormatCode>
<ImmediateDestinationName>TEST Destination </ImmediateDestinationName>
<ImmediateOriginName>TEST Origination </ImmediateOriginName>
<ReferenceCode> </ReferenceCode>
</FileHeaderRecord>
<Batch>
<BatchHeaderRecord>
<RecordTypeCode>5</RecordTypeCode>
<ServiceClassCode>200</ServiceC...
```

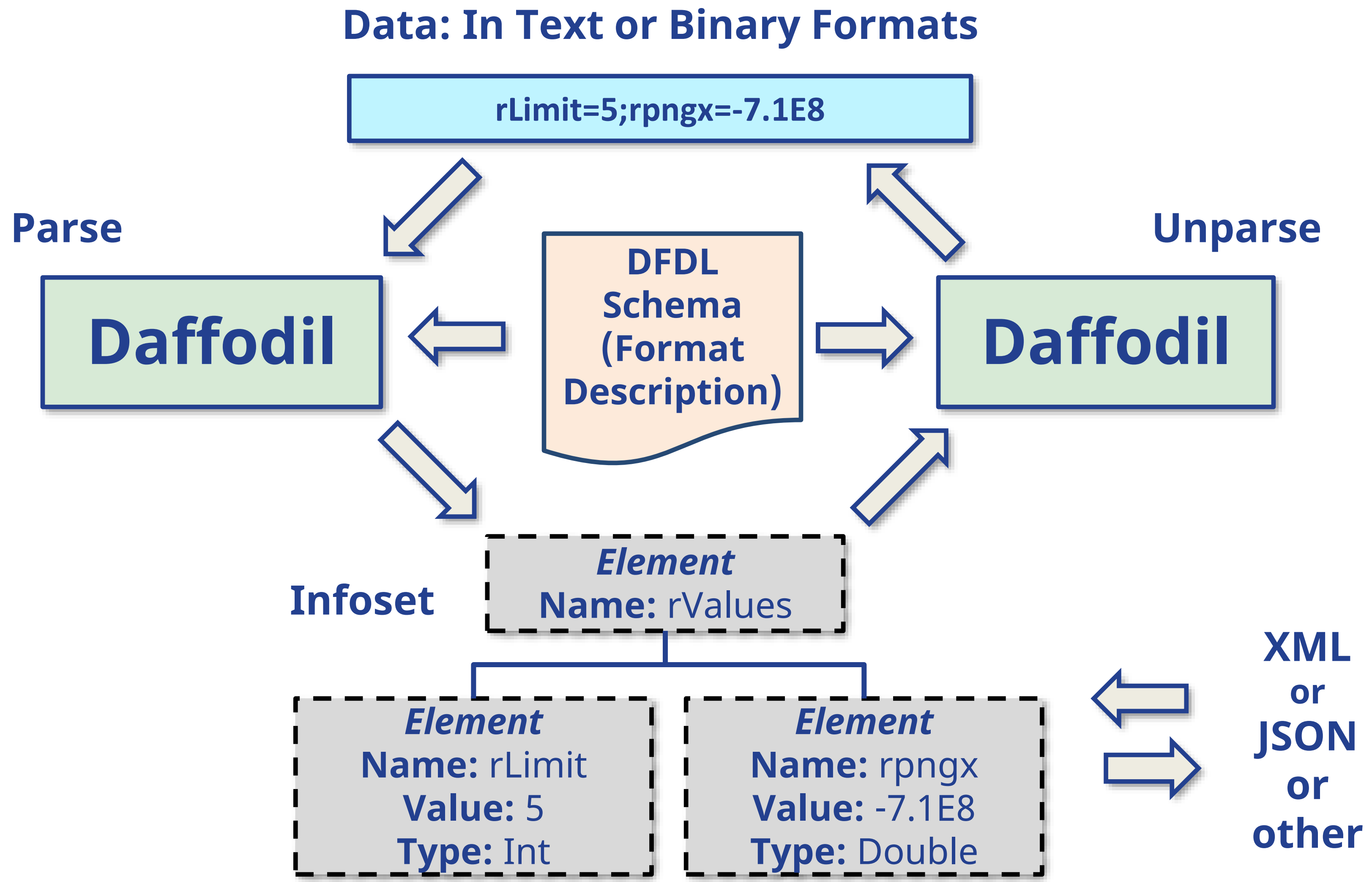
Prefer my NACHA as JSON Please...

```
{ "ACHFile":
{ "FileHeaderRecord":
{ "RecordTypeCode": "1",
"PriorityCode": "01",
"ImmediateDestination": " 123456789",
"ImmediateOrigin": " 987654321",
"FileCreationDate": "071030",
"FileCreationTime": "1634",
"FileIdModifier": "A",
"RecordSize": "094",
"BlockingFactor": "10",
"FormatCode": "1",
"ImmediateDestinationName": "TEST Destination ",
"ImmediateOriginName": "TEST Origination ", "ReferenceCode": " " },
"Batch": [ {
"BatchHeaderRecord": {
"RecordTypeCode": "5",
"ServiceClassCode": "200",
"CompanyName": "VIA LICENSING CO",
"CompanyDiscretionaryData": " ",
"CompanyIdentific...
```

DFDL Schemas

Public (github)	MIL-STD-2045	EDIFACT	iCalendar
	PCAP	IBM4690-TLOG	IMF
	NITF	ISO8583	SHP (shape file)
	PNG	BMP	KNXNet/IP(indust. control)
	JPEG	GIF	Siemens S7 (indust. control)
	NACHA	Praat TextGrid	Asterix (Cat 034, 048)
	VCard	ARINC429-PoC	MagVar
	QuasiXML	IPFIX	AFTN Flight Plan
	Geonames	Syslog	
FOUO / CUI (DI2E.net)	VMF		SOTF
	VMF_S2S unit-normalizing (Rev A)		JICD
	USMTF ATO (MIL-STD-6040)		NACT
	LINK16 (NATO STANAG 5516) - non-normalizaing		JREAP-C
	LINK16 (MIL-STD-6016F subset) - normalizing		DISV6
	A-GNOSC REMEDY		SIMPLE (STANAG 5602 Ed 3)
	ARMY DRRS		
	USCG UCOP		
	CEF-R1965		
	GMTIF (STANAG 4607)		
Commercial License $$$	SWIFT-MT (IBM)	USMTF ATO, ACO, etc. (Tresys)	
	HIPAA-5010 (IBM)	LINK16 (MIL-STD-6016 E, F, G) (Tresys)	
	HL7-2.7 (IBM)	VMF (MIL-STD-6017 A, B, C, D) (Tresys)	

Parse and "Unparse" Data



Daffodil Integrations

- XProc - Calabash
- Apache Spark (via XML)
- Apache NiFi
- Apache Storm
- SoftwareAG™ Integration Server (aka WebMethods™)
- Tresys products/services
- Other companies also!

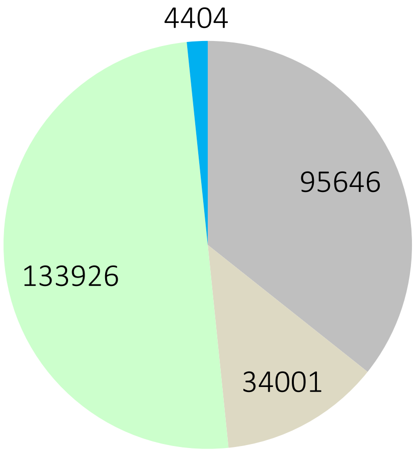
- Potential
  - Apache Tika
  - Other data-handling frameworks - Flink, Hadoop,....

Daffodil Coolness

- Library suitable for integration
- Daffodil Contains
  - Real compiler for DFDL schemas
  - Efficient low-level runtime for parse/unparse
- 100k+ lines of *Scala*

- We need developers motivated to ***kill the data format problem once and for all***.
- Java programmers who want to learn Scala wanted!

Lines of Code = 267K*



Cool Daffodil Ideas...

- Add support for pointer-based formats like TIFF and ZIP.
- SAX-style event streaming
- Directly construct your framework-native data representation with tight metadata coupling
  - E.g., Spark Struct objects
  - Removes the XML/JSON conversion overheads
- Ultra-fast small-footprint backend for Daffodil
  - Generate C/C++
  - Generate FPGA for wire-speed parse/unparse
- Data-Format Debugger
  - Graphical display shows data and parsed tree and schema interactively

END