TeamRoom Plus

IBM TeamRoo

This OPEN document will not be filed. It is being kept active.

Meeting about Meetings\OGF

Project	**DFDL 1.0**
Meeting Date	**29-Jan-13 (Tues)**
Meeting Time	**15:00 - 16:00**

Created by Steve Hanson on 09-Mar-11
Last Modified by Steve Hanson on 30-Jan-13

OGF DFDL Working Group Call , 29 January 2013

Agenda

Prepare for your meeting by describing the objectives (both immediate and long-term, if appropriate) of the meeting; and describe key plan details.

1. Daffodil Open Source Project
Status.

2. DFDL schemas web community
IBM has created a GitHub repository organization for DFDL schemas for commercial and scientific formats .
Not just for IBM authored schemas, anybody can contribute schemas that have reuse potential.
Link: https://github.com/DFDLSchemas

3. New glossary entries for character set encoding terminology
Review terms and definitions from Mike.

4. Clarify - minOccurs and maxOccurs on model groups
Spec section 5.1 says that these are not supported by DFDL.
But setting minOccurs="1" and/or maxOccurs="1" is the same as omitting them, so is that really an error?
If so it's one more thing that a DFDL user might have to change, if his starting point is an existing XML schema.

5. Use of xs: prefix when referring to XML Schema attributes such as 'fixed' and 'nillable'
The spec uses this prefix but technically that's not correct as the attributes are in the notarget namespace.

6. Use of terms 'representation property ' and 'format property'
Ideally should stick with one term. There is ambiguity with the former though because of the dfdl:representation property.

7. AOB

Minutes

Meeting Minutes

Reflect on your meeting as you record all topics and issues discussed , and any tabled conversations. What went well, or what would you do differently next time? Document those so others can take advantage of your learning .

Attendees
Steve Hanson
Tim Kimber
Mike Beckerle
Suman Kalia

Apologies

Minutes

1. Daffodil Open Source Project
Not discussed.

2. DFDL schemas web community
Git is a very powerful VCS and we need some recommendations from expert Git users as to the best way to use it with the new DFDL Schemas GitHub repositories. In particular, how to push changes to a repository, as there are several ways. Mike will talk to Tresys who have used Git a lot.
New **action 200** raised.

3. New glossary entries for character set encoding terminology
Reviewed terms and definitions from Mike. Discussed the correct definitions for DFDL of 'codepoint' and 'character code' which are not necessarily the same as the Unicode definitions. Mike to have a second pass through his definitions. New **action 201** raised.

4. Clarify - minOccurs and maxOccurs on model groups
Agreed that the intent behind spec section 5.1 when it says these are not supported by DFDL is that setting minOccurs="1" and/or maxOccurs="1" is allowed and is the same as omitting them.
Errata taken.

5. Use of xs: prefix when referring to XML Schema attributes such as 'fixed' and 'nillable'
The spec uses this prefix but technically that's not correct as the attributes are in the notarget namespace. Agreed to change the spec to use more general language such as 'XSD's xxx attribute'.
Errata taken.

6. Use of terms 'representation property ' and 'format property'
The latter is the general term for a DFDL property that can appear on a format annotation.
The former specifically excludes ref as it can not be scoped. But there are other properties that do not actually affect an object's 'representation' and can not be scoped - inputValueCalc, outputValueCalc, hiddenGroupRef, choiceBranchRef and elementId. So there really two kinds of property, how best to refer to them. Term 'representation property' is nice but there is the ambiguity with dfdl:representation. Didn't conclude, **action 202** raised.

7. OGF 37
Mike would like to attend and cover the soon-to-be published draft of the DFDL spec, the latest on the Daffodil parser, the latest on IBM DFDL, and the GitHub DFDL schemas organisation.
IBM will provide some customer references if possible. Mike will contact Alan Sill and submit a session.

Meeting closed
16:00 UK

Next call
Tues 5th February 15:00 UK

Create Action Items

Record the to-do's and individuals assigned by entering the appropriate information in the form below . Press the "Create Action Items" button to create specific to do's that can be tracked in the assignee's Work for Me views. " All Action Items will be tracked in the Action Items and Other Meeting Documents tab .

Action Items and Other Meeting Documents

Subject	Document Type	Created	Modified

Next action: **203**

Actions raised at this meeting

No	Action
200	**Establish recommended practices for pushing changes to GitHub (Mike)** 29/1: Mike will talk to Tresys who have used Git a lot .
201	**Define character set encoding terminology (Mike)** 29/1: Mike will update his draft proposal based on discussion .
202	**Term 'representation property ' (All)** 29/1: Two kinds of format property, how best to distinguish them in the spec.

Current Actions :

No	Action
123	**DFDL tutorial (Steve)** 13/10: Draft of first 3 chapters has been written and will be distributed to WG 10/11: Posted to grid forge here (http://forge.gridforum.org/sf/go/doc16106?nav=1), work continuing at IBM to define a standard example-based chapter framework and to author additional chapters. Contributors welcome! 17/11: Steve, Stephanie and Alan had a meeting to discuss the best structure for the tutorial

and decide which examples to use throughout. The meeting raised more questions. Further discussions will be held.
24:11: The list of topics to be covered in the remaining lessons has been produced and a lesson template. Alan will write lesson 4
01/12: Alan has started lesson 4 which covers fixed and variable fields and arrays.
08/12: Alan has almost completed lesson 4. Will send out for review.
15/12: First draft of lesson 4 is available for review. Alan to send to Bob and Joe.
22/12: Alan has distributed drafts for tutorials on Basic Structure and Optional/Repeating elements. Please review
12/01: Alan distributed a tutorial for choices and updated the others. Alan and Steve reviewed them and updated versions will be sent soon. Should start on the 'representation' tutorials soon.
19/01: The tutorials for basic structure, optional/arrays and choices have be updated. Please review. The tutorial for text elements should be available soon.
26/01: No comments received about 3 tuorials distributed last week. Alan is still working on Text representation.
02/02: Steve has sent comments on three tutorials. Alan to send updated versions by the end of the week. Alan has also distributed the first part of the tutorial on text representation and would like feedback.
09/02: Steve had reviewed tutorials 3,4,5 and updated versions have been distributed. Joe reviewed lesson on text elements.
Main points. Using 'represented as text' is confusing. Examples are too cluttered. Suggest simple targeted examples but still build up to final complete schema
23/02: New versions distributed and Steve has commented.
02/03: Alan has published the final versions of tutorials 4,5,6 and is working on text respresentations. There was some discussion about the detail that needs to be covered. Should limit it to 'common usage' and refer to the spec for details of edge cases.
09/03: Alan distributed an update to the text tutorial. Please review.
30/03: Steve has spent half a day tidying up lessons 1 to 6 and has uploaded them as pdfs to gridforge. They are now more coherent, and many inconsistencies and errors fixed. Ownership of draft lessons (text properties, binary properties, advanced features) has been passed to Steve. Also need to make a schema available for the examples.
13/04: Steve is working on the text properties tutorial.
04/05: No progress
...
15/06: This is on hold until Steve clears up spec issues and other workload. Steph has looked at the later lessons, and noted that they are more direct compared to the more wordy earlier lessons.
28/06: On hold.
...
29/11: Tim offered to take a look at the next outstanding tutorials. Steve / Tim to discuss
6/12: No progress
...
10/01: No progress, offer from Mike to help. First step is to make any corrections due to errata.
17/01: No progress
24/01: No update
31/01: Daffodil project team will be working their way through the existing tutorials and reviewing
14/02: Daffodil team to start reviewing tutorials hopefully this Friday.
21/02: Moved to this coming Friday
28/02: No update
13/03: No progress
21/03: No progress from Daffodil team. IBMers are starting to use the tutorial and will feedback any comments.
28/03: No change
05/04: Steve will send Alan's two draft lessons on binary & text data to Mike to complete.
17/04: No progress

	8/5: No update ... 4/9: No progress: 11/9: IBM DFDL infocenter will start to reference these directly before the end of the year, so they need updating soon. 18/9: Noted that several requests have been received asking for chapters 7 to 17 as implied by chapter 1. At minimum chapter 1 needs updating to make it clear what exists today. 28/9: Steve has updated and re-issued chapters 1 to 3. ... 29/1: No further progress
140	**Spec issue: Parsing: 'missing' v 'empty', role of initiators , default values (All)** 01/06: See minutes. 08/06: Still under discussion. Tim has sent Mike a selection of data formats to guide the discussion. 15/06: Not discussed - an extra call has been scheduled to go through this. 28/06: A series of extra calls are being held between Mike, Steve, Tim and Steph. 05/07: Next extra call is Wed 6th July - Steve to send invite 12/07: Two more calls held. Next call is Wed 13th July. 19/07: More calls held, next call is Fri 22nd July. 26/07: More calls held, good progress 16/08: Steve will set up next call for when Tim has returned from holiday 23/08: Two more calls scheduled for this week, remaining issues: separator suppression, empty strings, sparse arrays (see action 136) 30/08: Call held earlier today. Still remaining - separator suppression (matrix); sparse arrays; empty strings; empty value delimiter policy. Steve to summarise where we have got so far before remaining items are tackled. 20/09: Steve has summarised where we are with action 140, which Tim and Mike have reviewed. 2 hours call planned for Thursday. 27/09: Calls held, progressing the separator suppression behaviour 4/10: More calls held, progressing separator suppression, sparse arrays (see action 136) and emptyValueDelimiterPolicy behaviour 18/10: All issues now in a single document, call held earlier today. Next call Thursday. 1/11: Extra calls still ongoing 8/11: Extra calls still going, when action 140 document complete will send to Steph for review 15/11: Mike is verifying the action 140 conclusions by writing an algorithm in SCALA which can be ultimately be used in Daffodil 22/11: Call to be held next week 29/11: Next call 30/11 6/12: Next call 7/12 13/12: Had call earlier today, making good progress. Next call first week of Jan. 10/01: Next call to be scheduled for Wed or Fri 17/01: No call last week, call tomorrow 24/01: Two calls held, next call Wed, looking at separator suppression 31/01: Separator suppression discussions ongoing, proposal to rename policy and enums. Call later this week. 14/02: Now looking at separator suppression and unparsing. Next call on Friday 21/02: Separator suppression on parsing/unparsing matrix agreed. 28/02: Two more calls this week 13/03: Call tomorrow 21/03: Two more calls held 28/03: Hopefully all issues now addressed. Steve to start folding in action 140 document comments into the body of the document. 05/04: Still with Steve. Noted that action 140 will not be in the next spec rev, likely the one after. 17/04: No further progress 8/5: Steve has started rewriting the action 140 document

23/5: Steve continuing the work on the action 140 document. Empty/missing/defaults and Arrays have been rewritten. Separators not started. Please review.
12/6: Steve will resend the latest action 140 document for review. Note use of 'missing representation' to describe zero length input data with same semantic as missing altogether.
19/6: Latest action 140 resent. Separate call on Thursday this week.
26/6: Call held, revisions need to be made before action 140 can be circulated more widely. Next call will be Tues 3rd July.
3/7: Steve not found time to update action 140 doc, call postponed to 10th July
10/7: Call held, document still being refined. Spin off **action 179** to sort out use of term 'representation' in spec and grammar.
17/7: No progress on core action 140.
25/7: No progress - Steve will aim to finish refinement before Aug 16
31/7: No progress
7/8: No progress
14/8: Steve has started to create v16 of the action 140 document and will mail it to Mike & Tim this week
4/9: v16 mailed to Mike & Tim - some issues noted in the email - Steve will set up a separate call
11/9: Separate call held. Re-examining the role of minOccurs for occursCountKind 'parsed', 'expression', 'stopValue'. For these occursCountKinds, where the occurrences are extracted without reference to minOccurs, it seems more natural that a minOccurs violation is not a processing error but instead it is just a validation error. Need to evaluate the knock-on effect of this proposal - it potentially affects points of uncertainty, default values, use of terms required & optional, and separator suppression.
18/9: Tim and Steve have worked through the proposal, it looks good in principle. Steve to update action 140 document and see if anything problematic surfaces.
Note that current definitions of 'required' and 'optional' retained - it's just that now a required element missing from the infoset (after defaulting applied) is not necessarily a processing error (that now depends on occursCountKind). Please review for next call.
28/9: Tim & Mike to review Steve's updated document. Other things to do before it can be incorporated into the spec:
- SMH1 comment: What do dfdl:xxxlength() functions return when rep is absent? Error or 0?
- English words for separator suppression tables
- Decide the fate of Appendix A? Perhaps replaced by tutorials?
2/10: Key to explaining all this when it is rolled into the spec is defining the different reps plus 'missing' plus concepts of 'well-formed' and 'badly-formed' in the glossary. Clearly sections 13.15, 14 and 16 are affected in a major way, but it is likely that it affects several other sections such as 9.
Discussed the dfdl:xxxLength() functions and what they should return when there is nothing in the infoset. This led to a more general discussion of whether a failure to find a path should be treated as a schema definition error or a processing error. New **action 188** raised.
16/10: Steve to take one more pass through the document in the the light of the above, and also try to put into words the separator suppression tables. In order to do this need action 187 needs resolving.
23/10: Action 187 resolved far enough to allow Steve to proceed with document pass.
30/10: Still with Steve.
5/11: Still with Steve.
12/11: Steve has updated action 140 to v018 to reflect action 187 terminology, which has made it clearer. Next step is to create readable descriptions of the separator suppression tables.
20/11: No further progress. Consider making separator suppression stuff a separate action?
27/11: No progress
4/12: Steve has started work on the separator suppression policy descriptions. Key to this is a definition of 'potentially trailing' as this is one of the criteria that allows separator suppression to take place. Steve will circulate a draft definition as it is not as obvious as it appears. Also it is important that the WG re-review action 140 v018 as Steve believes there are inconsistencies in there.

	A call may be scheduled later in the week if needed. 11/12. Steve sent out a v019 of the action 140 document which corrected some mistakes from earlier updates, added comments for issues that needed resolving, and contained a rewritten section 4 on separator suppression. An extra call was held on 7/12 which addressed the comments and reviewed section 4. Actions arising were: - Mike to reword some of the section 4 text to make clear the distinction between a potentially trailing element and actually trailing occurrences of an element. - Steve to correct a mistake in the separator suppression tables where column headings got swapped and resend the spreadsheet - Tim to create some introductory words to explain the motivation for the separator suppression property 8/1: v020 and v021 of the action 140 document have been circulated. Steve has spotted a possible error in always equating 'empty' representation to 'known to exist', after reading Mike's choice example email, so we need to revisit that. Tim has created words for separator suppression introduction. Review and comment please. 15/1: Tim's separator suppression intro approved. Discussed Mike's example of a choice with a complex element branch that evaluated to empty representation. This is an instance of the behaviour of a required complex element when it has empty representation. Action 140 currently says that defaults are not applied but does not say whether this is a processing error or you get a child-less element in the infoset. Steve had expected the former, but is not happy that this is appropriate for all circumstances. Steve will think about all the instances of empty representation, taking into account initiator/terminator, lengthKind, emptyValueDelimiterPolicy, etc. 22/1: Reviewed the action 140 behaviour for all use cases of empty representation. Essentially there are three main cases, 1) simple element (non-string), 2) simple element (string) and 3) complex element, each of which has sub-cases for optional occurrence versus required occurrence, and there is also EVDP to take into account. Reached a proposal where all were comfortable. Steve to write up and circulate for review. 29/1: Reviewed Steve's proposal, looks good. Steve will update action 140 document. Hopefully this can be merged into the next draft of the spec (ie, post the draft that includes errata v011). The action 140 changes should be merged on their own for ease of reviewing.
172	**Clarify how a DFDL string literal is matched against the data stream (Tim)** 23/5: Non-trivial algorithm, worth stating it in the spec. .. 25/7: No progress. 31/7: Tim has been making notes but nothing written up formally. Will include treatment of %WSP*; ... 29/1: No further progress
193	**Prescribe a single regex language or subset (Steve)** 12/11: Created 20/11: Not discussed 27/11: IBM DFDL intending to use Java regex for Java, and ICU C regex for C. Steve to contact ICU to find out any behavioural differences between ICU and Java. Need to decide default behaviour for configurable options like whether the wildcard character '.' includes line end characters, and case sensitivity, etc. 4/12: Steve has raise ICU ticket to document Java and ICU4C differences, awaiting response. 11/12: Acknowledgement from ICU but not a full response as yet 8/1: Still awaiting a full response. Steve will chase. 16/1: ICU have created a spreadsheet listing differences, there are only a handful. IBM team is analysing. Query over wording in 12.3.5 that says "The DFDL processor scans the data stream to determine a string value that is the longest match to a regular expression." The pattern itself dictates greediness so the word 'longest' is not needed. **Errata taken.**

	Noted that the link to PERL regular expressions in section 30 is incorrect. **Errata taken** Need to decide default behaviour for configurable options like whether the wildcard character '.' includes line end characters, and case sensitivity, etc. A scenario was discussed where a text element with lengthKind 'pattern' might be followed by a binary element. In general this is not advisable, you have to be sure that the binary data can not start with something that matches the pattern. Assuming that to be so, the pattern could be such that it caused an encoding error when trying to match. Can use the new errorPolicy property (errata 3.13) to use substitution character instead of error. This action with IBM to make a proposal. 22/1: ICU C has some extra capability over Java 6 especially in the area of Unicode matching, some of which are fixed in Java 7. Feeling was to go with the ICU C / Java 7 level of support. IBM to sent ICU spreadsheet to Mike and to write down other differences. ICU C has a time limit on the processing time for a regex. Agreed that implementations can offer this if desired. 29/1: IBM has sent ICU spreadsheet and a summary of differences, for use as basis of errata. Discussed whether the differences should be flagged as errors or warnings. Agreed that \w is a warning and the rest are errors. IBM to revise the errata.
197	**Publish Errata v 011 and fold into DFDL spec (Steve/Mike)** 20/12: Steve has a draft of Errata v011 for review. Writing up the errata has revealed a clash between 2.95 (assumption of byte alignment for delimiters) and 2.107 (allow 7-bit ASCII encoding) so need to rectify this. 27/12: Corrections made to Errata v011 and resent for review. Mike responded with comments. One comment led to some discussion, namely changes to errata 3.13 regarding replacement characters. Should encountering a decoding error when parsing variable length xs:strings be an error or use the substitution character (like would be done for specified length xs:strings) ? 4/12: Given the imminent Christmas break it looks like incorporating errata v 011 into the spec might not take place until January. If so then hopefully action 140 might make it as well. Regarding the decoding issue Mike will have a go at rewording errata 3.13. 11/12: Mike has rewritten errata 3.13. Proving difficult to find a sensible split between when to use substitution character and when to give a processing error. Need to revisit the requirements from consumers before deciding. Steve to check IBM guidelines. 8/1: IBM has some globalization guidelines that imply a) fallback mappings are considered data corruption unless you're just displaying the data to a screen, and b) data corruption is not allowed. That argues for making encode/decode failures into DFDL processing errors. But IBM's older MRM and WTX technologies use substitution characters, so some users appear happy with that approach. This is leading us back to an error/replace property as per the original errata, if so then what exactly is its placement and scope? 16/1: Mike will has reworded the proposal for 3.13, which is now a lot simpler. Reviewed it and resolved issues. Steve will send Mike the latest errata v011 document (there have been a couple of editorial corrections). Mike will update errata 3.13 and 2.95. The errata v011 document will then ready to be incorporated into the spec. 22/1: Updated errata v011 sent for review, Steve has commented. Mike has started to fold into DFDL spec. 29/1: Mike sent updated errata v011 and updated DFDL spec that incorporates it, to Steve for review. Steve addressed the outstanding errata v011 issues, sent back to Mike, and will post to Redmine.
199	**Review the unordered sequence rewrite -into-a-repeating-choice section (Tim)** 16/1: This needs to be worded in a way that allows implementations to check for occurs violations in a manner which is consistent with the description of array processing in action 140 document, and is not overly constraining on implementers. 22/1: Tim will propose revised words as he proceeds with the IBM implementation

	29/1: No progress

Closed actions

No	Action

Deferred actions

No	Action
129	**Press release to publicise DFDL (Steve)** Steve is pulling together a press release at IBM. Want to include as many contributors and interested parties as possible.NCSA are keen to be included. Also likely that US National Archive will want to be included. Mike has indicated OCO are too. 17/11: no progress ... 08/12: Still no response from IBM press office 15/12: no progress .... 09/03: No progress 30/03: Making this action deferred until IBM is in a position to say something more concrete about any implementation.
131	**Transformation of DFDL properties to a canonical form (Joe)** 08/12: Joe has produced a XSLT to transform a DFDL schema to a canonical element form. When tested it should be made available on the WG gidforge site. 15/12: Alan tested against test dfdl schema which worked correctly (after fixing some errors in the schema) 22/12: no update 12/01: Joe has some defects to fix before making available on gridforge. 19/01: There is a difficult problem to solve before Joe make the style sheet public 26/01: Working on problems 02/02: no progress 09/02: As it wasn't a simple as exoected this will be treated as a low priority action 23/02: Low prioity 09/03: Low priority 30/03: Deferring for now
066	**Investigate format for defining test cases (All)** 25/11:IBM to see if it is possible to publish its test case format. 04/12: no update ... 17/02: IBM is willing in principle to publish the test case format and some of the test cases . May need some time to build a 'compliance suite' 24/03: No progress 03/03: Discussions have been taking place on the subset of tests that will be provided. 10/03: work is progressing 17/03: work is progressing 31/03: work is progressing 14/04: And XML test case format has been defined and is being tested. 21/04. Schema for TDML defined. Need to define how this and the test cases will be made public 05/05: Work still progressing 12/05: Work still progressing 02/06: Work still progressing on technical and legal considerations ... 25/08: Will chase to allow Daffodil access to test cases. The WG should define how implementation confirm that they 'conform to DFDL v1' 01/09: IBM still progressing the legal aspect. Intends to publish 100 or so tests as soon as it can,

	ahead of a full compliance suite. 08/09: IBM still progressing 15/09: IBM still progressing, expect tests to be available within a few weeks 22/09: IBM still progressing, expect tests to be available within a few weeks 29/09:Test cases are being prepared. 06/10: Some test cases should be available next week. Steve would like to be able to show the test case information at OGF 30. 13/10: Still progressing 10/11: Legal issues cleared, IBM in process of collecting 100 example test cases, ideally ones that fit the 'extended conformance' of NCSA Daffodil 17/11: Work is progressing on verifying the test cases. It should be possible to distribute to the WG in 2 weeks. 24/11: About half the test cases have been completed and are being reviewed internally. 01/12: Test cases should be available shortly 08/12: The test cases are in internal IBM review. Probably need a bit of reorganising before publication Stephanie gave a brief overview of the format of the test cases. 15/12: Ruth joined the call to provide the latest status. The test cases have been updated and a draft read.me produced. Although not ready for public distribution Ruth will send them to Joe for feedback. 22/12: Test cases were sent to Joe for initial testing which found some problems in the Daffodil parser 12/01: All current tests use a default format whih Daffodil doesn't currently support. Joe suggested that there should be test that defined the same function using different definition forms. Also suggested that default formats should be provided by the WG. This had always been the intention. Action 133 raised to track. 19/01: There is currently no resource available in IBM to make more tests available. IBM to discuss how/if it can make a 'minimal compliance test suite' available. 26/01: Action kicked off within IBM. There was a brief discussion abot naming and organisation of test cases but no preferences were expressed 02/02: IBM will not have the resources to develop a full test suite in the near future. Steve suggested that we produce a list of required test cases so that anyone could supply them. 09/02: Steve had previously sent a list of areas to be tested. Please review. 23/02: Please review Steve's list of areas to be tested 02/03: Alan had reviewed Steve's list and we went through his comments. Agreed there is no need for separate tests for the infoset or for dfdl: property lists, unions etc but comment will be added that these should be exercised during property testing. 09/03: Alan updated the test document. Need more introduction and perhaps adopting the OGF template. 30/03. Ownership of test document passed to Steve. This action is merged with 112 and will cover all aspects of compliance suite. 13/04: IBM will not have time to create a compliance suite in the near future. Probably best to make this action deferred for now. ... 10/07/2012: Discussed schemes to create interchangeable tests. Ideally need a DFDL defined error code per failure, in conjunction with specific inserts.

Work items:

No	Item	Owner	Target	Status
043	Track errata list for 1.0 of the spec.	Steve	N/A	Draft 011 on grid forge.
044	Incorporate errata list into DFDL spec.	Steve/Mike	N/A	Draft merged document.