HR001118S0054
Safe Documents (SafeDocs)
Frequently Asked Questions

As of October 5, 2018

Q57: How much detail is needed in the Excel file "BAA LoE Template", i.e., do we need to report subtasks?
A57: It Provide sufficient detail so a logical mapping between proposed technical content and tasks allows us to accurately understand how you have broken down the problem and LoE involved for each of the proposed tasks.

Q56: For the Optional Phase 3 phase do consultants need to be listed and priced?
A56: For each phase, proposals should provide sufficient detail so that the proposed technical content and tasks map to and justify the proposed personnel and/or types of personnel needed and their costing.

As of October 4, 2018

Q55: Do subs need to submit proposals to DARPA for their proposed tasks under the BAA via the DARPA BAA Submission Website?
A55: It is the prime proposer's responsibility to transmit the complete proposal to DARPA via either the DARPA BAA Submission Website or Grants.gov. As stated in the BAA:
"The proposer is responsible for the compilation and submission of all subcontractor/consultant cost proposals. At a minimum, the submitted cost volume must contain a copy of each subcontractor or consultant non-proprietary cost proposal (i.e., cost proposals that do not contain proprietary pricing information such as rates, factors, etc.). Proprietary subcontractor/consultant cost proposals may be included as part of Volume 2. Proposal submissions will not be considered complete unless the Government has received all subcontractor/consultant cost proposals."
The BAA (p.40) also addresses the submission of proprietary subcontractor/consultant cost details via email.

Q54: Should the table of individual time commitments in the proposal include commitments on non-DARPA projects?
A54: Yes. For key personnel identified in the proposal, all time commitments should be listed. However, entries for non-DARPA commitments need not be detailed. See the example provided on p.35-36 of the BAA.

Q53: Can a TA3 performer participate in open industry forums and association discussions related to SafeDocs?
A53: Yes. However, the TA3 performer must not discuss any aspects of the test corpora that will be used in the SafeDocs evaluation, and must safeguard the confidentiality and integrity of the evaluation. The TA3 performer may discuss reference corpora that have been made available to all performers.

Q52: Is there a specific forum for industry organizations to advise DARPA of their capabilities related to the SafeDocs program?
A52: Once the program has gotten underway, following the scheduled kick-off meeting, DARPA welcomes future ideas that may be of relevance to SafeDocs, as well as to the thrust areas of the Information Innovation Office (i2O). Such ideas can be conveyed in white papers and abstracts, which can be submitted to the current office-wide I2O BAA.

Q51: Does a subcontractor/consultant require a Taxpayer Identification Number (TIN) or Commercial and Government Entity (CAGE) code to perform work on the project?
A51: DARPA does not require a subcontractor or consultant to obtain TIN or CAGE code.

Q50: Does submitting an abstract or proposal require a TIN or CAGE code?
A50: No. A TIN or CAGE code is not required to submit an abstract or proposal. However, to receive an award under a specific program, a proposer must have both in place, as well as other award eligibility requirements, prior to award. See A48.

Q49. Do foreign companies/entities require a CAGE code to perform work on the project?
A49: If proposing as a prime, you will need to follow the requirements outlined within the BAA for submission. Foreign companies should look at: https://fsd.gov/fsd-gov/answer.do?sysparm_kbid=699953826fcc710045b164826e3ee455&sysparm_search for further information. As stated previously, to perform work under a specific program, a proposer must have a CAGE code in place prior to award.

Q48: Is it necessary for a prime proposer to have a CAGE code for proposal submission?
A48: A prime proposal can be submitted against the SafeDocs BAA if the proposer doesn't currently comply with all of the necessary award eligibility requirements, such as not having a CAGE code, as long as the submitter is in the process of obtaining one. Proposers should document in their proposal where they are in the process of meeting such an eligibility requirement. If selected for a full or partial award under SafeDocs, that award would be contingent on the proposer completing and meeting all of the eligibility requirements, such as successfully obtaining a CAGE code.

As of October 1, 2018

Q47: On page 23, para 2 under "Exercises" the BAA specifies that DARPA will provide Government subject matter experts (SMEs). Can a performer on TAs other than TA3 also serve a Government SME?
A47: The Government SMEs are not allowed to propose under the BAA in order to eliminate any conflicts of interest. The PM intends for the program to be structured in a manner that allows the insights of format experts working on TA4 to be communicated to the program at large, and will facilitate sharing of relevant information via both the ACA and the contracted deliverables. Amendment 1 to the BAA that clarifies this point has been published to FedBizOpps and Grants.gov.

Q46: What is the level of technical or costing detail required for proposed plans of industry engagement?
A46: DARPA does not direct the content of proposals. DARPA expects the strongest proposals and proposers should take every opportunity to strengthen their proposals with relevant technical and planning detail, subject to the page limit.

As of September 11, 2018

Q45: Does Sec. C.2 of the BAA imply that SafeDocs encourages development of recursive descent parsers rather than table-driven parsers?
A45: The BAA is not prescriptive with respect to any parsing technology. Rather, it poses requirements for the desired developer and auditor experience with the proposed solutions.

Q44: Can a proposal succeed by addressing only parts of a TA, e.g., by focusing on validating documents but not streams, or by focusing on one development model?
A44: No. Strong proposals must address all requirements of the BAA for the particular Technical Area. Proposers are encouraged to form teams to create strong proposals. Proposers can use the teaming site is available at https://www.schafertmd.com/darpa/i2o/SafeDocs/pd/?p=teaming. See also Q6.

Q43: Will the program's TA1 goals be fulfilled by producing a set of prescriptive solutions for resolving the inherent ambiguities of commonly used formats such as PDF, DOCX, ZIP, etc.?
A43: No. The program aims to develop methodology and tools applicable to a broad range of formats, to enable organizations to resolve ambiguities in formats in a theoretically solid and efficient manner. Specific safe unambiguous format subsets will be a result of this methodology.

Q42: Is proving the safety of arbitrary semantic actions that encompass the application's business logic in scope?
A42: No. Semantic actions in the SafeDocs scope are understood as a means to assure syntactic or structural properties of input data, not as a means of implementing the program's business logic. See also Q4.

Q41: May the proposals assume the existence of any trusted information about the document, such as its provenance, trusted out-of-band metadata, etc.?
A41: No. Any such assumptions are out of scope. All of the input is to be regarded as hostile.

Q40: Will the SafeDocs high-assurance translator pass the documents using the reduced safe subset unchanged?
A40: Yes. This is a reasonable assumption.

Q39: Is using the input validity theories developed in TA1 or TA2 to evaluate legacy code in scope?
A39: Yes. As long as these address the requirements of the BAA, it is in scope.

Q38: Is software that handles electronic documents has functionality that is not used in the wild and not in the official format specification, can this functionality be excised from the safe format?
A38: This decision will depend on the input validity theory approach of the proposal. Generally speaking, unused features are still likely to place additional load on the verification effort for the software, and may thus be prudent to avoid.

Q37: Is helping an enterprise organization to create their own secure formats and parsers in scope?
A37: Yes. A primary expected outcome of the SafeDocs program will be the methodology that an organization can efficiently apply to create its own safe formats and high assurance parsers for these formats.

Q36: Are logic-based approaches in scope?
A36: Logic-based approaches can be used to address the BAA requirements.

Q35: Are machine learning/data science approaches in scope?
A35: Machine learning/data science approaches can be used to address the BAA requirements. For example, automated processing of corpora of extant documents, as one of the sources of ground truth about the format, is in scope.

Q34: Can non-U.S. proposers apply?
A34: Yes. The SafeDocs program is considered fundamental research. Please see the BAA for further information.

Q33: Does the "Out of Scope" applied to cryptographic techniques apply to all TAs or might these techniques be permitted as part of designing documentation or implementation of non-crypto parsers?
A33: The BAA stipulates that no cryptographic protections can be assumed on the input data.

Q32: Is TA1 constrained to use only the datasets provided by TA3?
A32: TA1 is encouraged to develop their own datasets for their development as outlined within the BAA.

Q31: Is Natural Language Processing (NLP) in scope?
A31: NLP techniques can be used to address the BAA requirements. For example, automated processing of the natural language format descriptions, as one of the sources of ground truth about the format, is in scope.

Q30: Is teaming encouraged?
A30: Teaming is encouraged to create robust teams that will bring subject matter experts to the program to address the technical challenges. The proposers are also encouraged to engage with experts in security phenomena of extant data formats. The teaming website is available at https://www.schafertmd.com/darpa/i2o/SafeDocs/pd/?p=teaming.

Q29: What assumptions can be made on the input data to the system?
A29: Assume any input is hostile and needs to be safely handled. In particular, assume that no prior filtering on the input, no protective access control, and no cryptographic integrity protections are available.

Q28: The Proposers Day diagrams show the parser capability as fixing bad data. Is this correct versus predictably parsing subsetted constructs?
A28: The BAA distinguishes between high-assurance translators of legacy documents to the safe subset of the format, which must rely on expert judgments to resolve inherent ambiguities of a format, and verified parsers that handle the unambiguous safe subset. Translators must safely transform legacy documents to the safe subset, necessarily incorporating expert interpretations of the format's ambiguities. Verified parsers implement the safe subset specification, which is by design non-ambiguous. Please refer to the BAA for further discussion of the TA requirements and of the pitfalls of ad hoc fixing of malformed documents.

Q27: What estimated Kickoff date should we use in preparing budgets:
A27: We anticipate an April timeframe. See Q3.

Q26: Can TA1 and TA2 also do TA3 or TA4?
A26: TA3 is independent but is expected to work closely with TA1 and TA2. See Q23.

Q25: Where do you see work around files used in embedded technology (such as configuration files, boot files, partition tables, DEM certificates, etc.) existing. Further, is this type of files with in the field of interest?
A25: Complex binary file formats have a history of parsing vulnerabilities. As such, they are in scope.

Q24: How do you envision accounting for continued evolution of extant types?
A24: Methodologies and tools developed for comprehension of extant formats should be designed to adapt to changes in the formats, and to additions to the reference corpora.

Q23: Can an organization prime on an exclusive TA and form as a subcontractor for another exclusive TA (i.e., Prime TA1 and Subcontract on TA3)?
A23: No. There are collaborations expected between TA1 and TA3, however.

Q22: How likely do you believe the various standards committees are to adopt the simplified, safer Syntax?
A22: The theories and insights developed by the program should provide such committees with scientific grounds for adoption of SafeDocs-developed syntax. The specifications and theories will be open sourced to further facilitate adoption.

Q21: Do you envision different parser construction kits per data type?
A21: We are looking for general approaches applicable to broad ranges of data types.

Q20: Who will define a "Format's essential functionality" (Item B) near end of Summary in BAA?
A20: For evaluation purposes, TA3 and TA4 in collaboration.

Q19: Is there any prioritization or relative importance across data format? Or conversely, are approaches limited data format acceptable? (just like documents and not streams)
A19: We are looking to general approaches applicable to broad ranges of formats.

Q18: This is highly multidisciplinary! Will the project assist with teaming?
A18: Teaming is encouraged. There is a teaming website available as part of the registration page, https://www.schafertmd.com/darpa/i2o/SafeDocs/pd/?p=teaming.

Q17: How do we distinguish "essential" behavior from behaviors that are non-essential or can be "pruned"?
A17: Understanding of essential and intended behaviors that can be effectively secured will be a part of research in TA1. For evaluations, TA3 and TA4 will pose specific requirements for each exercise, working towards the maximum theoretically possible.

Q16: Is it acceptable for a TA1-4 proposal to focus on DOCX?
A16: We encourage general approaches applicable to a wide range of formats. Experts in specific formats are encouraged to make use of the teaming website, https://www.schafertmd.com/darpa/i2o/SafeDocs/pd/?p=teaming to engage with other proposers.

Q15: In TA2, is there any recommendation/spec on programming language?
A15: No. However, proposed methodologies and tools for creation of secure parsers must be accessible to industry programmers, as per the BAA.

Q14: Does the requirement of predictable execution given an input imply that the entire program must be verifiable, not just the parser? If no, is the predictable execution of the remainder of the program out of scope?
A14: We see verification of the parser as a necessary step to enable verification of full programs. We do not postulate that programs can be verified in their entirety, but rather that verified parsers should assure the pre-conditions relied on by the rest of the program.

Q13: Are the container formats in PDF evaluated as separate formats, i.e., GIF/JPEG/TIFF inside a pdf?
A13: Sub-formats and nested formats are in scope when essential to the format's intended function.

Q12: How do you envision interactions with domain experts taking place most productively in all technical areas?
A12: Strong proposals will outline plans to find and engage domain experts, and develop tools to use their time most efficiently.

Q11: Do the terms "formal" and "verified" refer to the mechanically verified proofs?
A11: Yes.

Q10: In a "constructive theory of security for parsers" what would some desired theorems be?
A10: Theory efforts of TA1 and TA2 are expected to answer this research question.

Q9: Will each engagement move to a new format, or will they pursue one with increasing depth? (Breadth vs Depth)
A9: A combination of these approaches will be used in evaluation.

Q8: Table 2 in the BAA refers to PDF. Is that notional, or is PDF being singled out as a required format to be addressed?
A8: PDF is going to be used in evaluations, but the program aims to solve the more general problem of which PDF is an example.

Q7: Can one organization submit two abstracts or proposals for one TA? Can one organization bid as a sub on two proposals with different primes for one TA?
A7: Yes.

Q6: Will proposals be considered to be monolithic, or might DARPA slice and dice among them?
A6: DARPA will reserve the right to select a full proposal, a part of a proposal (or none).

Q5: What is a DDS streaming data format? Please provide an example.
A5: DDS (Digital Data Stream) is a publish-subscribe standard that encompasses different kinds of machine-to-machine data exchanges, including a variety of streaming data formats, including audio and video.

Q4: Can we assume that semantic actions are "safe" or could they have arbitrary complexity?
A4: Proposed methodology must assure safety of semantic actions, up to a reasonable complexity, within possible theoretical limits.

Q3: What are the timeframes for funding the various technical areas?
A3: The estimated kick-off date is anticipated in April 2019, the TAs will run concurrently (as describe in the BAA).

Q2: What are the approximate levels of funding for the various TAs?
A2: As stated in the BAA, the level of funding for individual awards made under this solicitation has not been predetermined and will depend on the quality of the proposals received and the availability of funds.

Q1: Do you expect TA1 and TA2 teams to address "correctness" of a transformer from "wild" to "safe subset" to include the property "the document displays the same"? If so, how is the responsibility split between TA1 and TA2?
A1: Translators will necessarily incorporate heuristic judgments when the de facto format is ambiguous. The verified parsers for the reduced safe syntax will have no such limitation. See also Q28.