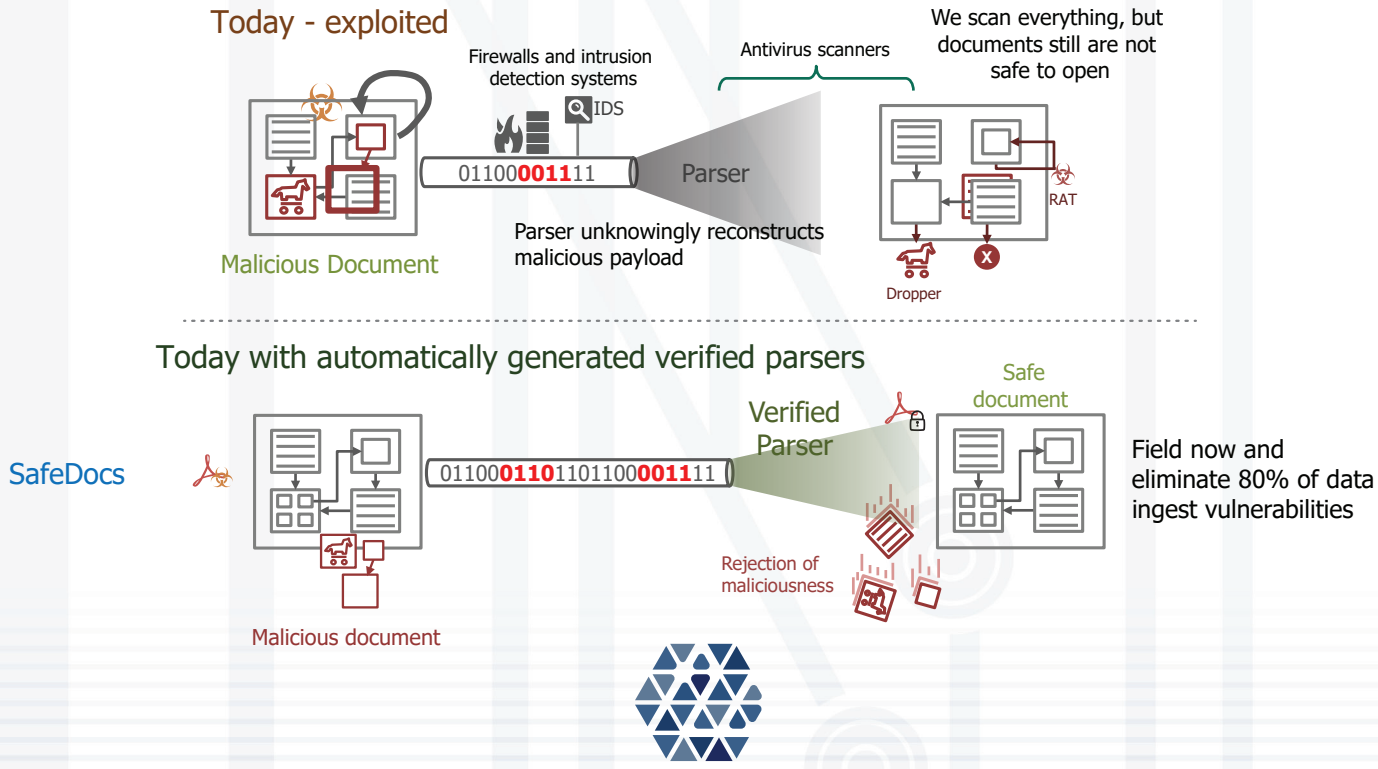
DARPA / I2O
SafeDocuments (SafeDocs) Program

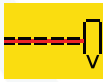
DARPA imagines a world without software vulnerabilities. SafeDocs and other I2O programs have been working to implement this vision through the development of formal methods for nearly a decade. Formal methods are mathematically rigorous techniques for the specification, development, analysis, and verification of software. The use of formal methods is motivated by the observation that, as in other engineering disciplines, performing appropriate mathematical analysis contributes to reliability and robustness. Nearly a decade ago, DARPA's High Assurance Cyber for Military Systems (HACMS) program demonstrated, on a DoD-relevant platform, that formal methods could be used to prove the system does what it is supposed to do, and only what it was supposed to do. Parsers are the code that converts serialized inputs, such as documents and streaming electronic data, into in-memory data. HACMS found that parsers account for a significant fraction of software vulnerabilities, which in turn leads to exploits that may cause mission failure. To reduce risk to the mission, electronic data formats must be modeled down to the wire format level, and the parsers that ingest and validate the data should be automatically generated from these models and formally verified. This was the goal of DARPA's Safe Documents (SafeDocs) program. SafeDocs researchers developed new methods and tools to allow people to trust what they see on their screens and to click confidently on documents. SafeDocs research and development resulted in reducing documents' complexity and using formally verified parsers to radically improve software's ability to reject invalid and malicious data, without impacting the key functionality of new and existing electronic data formats.



SafeDocs Open-source tools are changing the landscape

Tools for describing data formats and auto-generating parsing code:

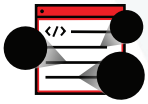
DaeDaLus: Data description language for defining data formats and generating memorysafe parsers in a variety of languages



Hammer: Declarative secure parser and scanner construction kit in C

Parsley: Declarative data format definition language that combines grammars and constraints in a modular way

Tools to understand behavior of existing parser code:

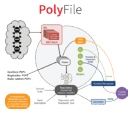


PolyTracker: Intelligent tracing of parsers written in C/C++

Graphtage: Command-line utility and underlying library for semantically comparing and merging tree-like structures

Tools for understanding document collections and format rules:

PolyFile: Exploring polyglot and "schizophrenic" file phenomena



File Observatory: System to enable visualization, search, and discovery of complex file format patterns and data

Format Analysis Workbench: Platform for running and analyzing the output from parsers dealing with a single file or streaming format

Tools to secure Python's data format used overwhelmingly in AI research:

Fickling: Decompiler, static analyzer, and bytecode rewriter for Python pickle object serializations

Links:

DARPA Resilient Software Systems Demo Day videos
https://www.youtube.com/playlist?list=PL6wMum5UsYvZhEOoP4YtAwtZdLSBIAltk

SafeDocs Tool Catalog
https://creative.spa.com/?s=safedocs-catalog-1&77dbe309

Best Practices for Secure Data Intake, Data Modeling, and Data Design paper
https://www.darpa.mil/sites/default/files/attachment/2025-06/best-practices-secure-data-intake-sergey-bratus-darpa-i2o.pdf

Resilients Software Systems
Stephen Kuhn
stephen.kuhn@darpa.mil

SafeDocs Program Manager
Dan Wallach
daniel.wallach@darpa.mil