REQcollect: Requirements Collection, Project Matching and Technology Transition

Luanne Goldrich
Stephen Hamer
Martha McNeil
Thomas Longstaff
Johns Hopkins University Applied Physics Lab
11100 Johns Hopkins Road
Laurel, MD 20723
luanne.goldrich, stephen.hamer,
martha.mcneil, thomas.longstaff@jhuapl.edu

Robert Gatlin
Department of Homeland Security*
4601 Fairfax Drive
Arlington, VA 22203
robert.gatlin@hq.dhs.gov

*The opinions expressed in this article are the author's and do not necessarily represent the position of the United States, the Department of Homeland Security, or any other entity

Emmanuel Bello-Ogunu
University of North Carolina, Charlotte
309 Bald Eagle Court
Moncks Corner, SC 29461
ebelloog@uncc.edu

ABSTRACT

This paper describes the evolution of REQcollect (*REQuirements Collection*). REQcollect was developed through several iterations of agile development and the transition of other projects. Multiple federal agencies have sponsored the work as well as transitioned the technologies into use. The parents of REQcollect are REQdb (*REQuirements Database*) and DART3 (*Department of Homeland Security Assistant for R&D Tracking and Technology Transfer*) [1]. DART3 was developed from three other projects: TPAM (*Transition Planning and Assessment Model*) [2], GNOSIS (*Global Network Operations Survey and Information Sharing*) [3,4] Aqueduct [5], a semantic MediaWiki extension.

REQcollect combines the best components of these previous systems: a requirements elicitation and collection tool and a Google-like matching algorithm to identify potential transitions of R&D projects that match requirements.

1. INTRODUCTION

Cybersecurity systems are comprised of a wide variety of interconnected systems. If these individual systems are not sufficiently designed and implemented to work well when integrated, opportunities for adversaries to hide and attack are created. To better enable researchers to broaden their scope and work together, they must be able to explore research in other areas and work across disciplines to create standards for communication and systems that work synergistically.

This paper describes the evolution of REQcollect (*REQuirements Collection*). REQcollect was developed through several iterations of agile development and the transition of other projects. Multiple federal agencies have sponsored the work as well as transitioned the technologies into use. The parents of REQcollect are DART3 (*Department of Homeland Security Assistant for R&D Tracking and Technology Transfer*) [1] and REQdb (*REQuirements Database*). DART3 was developed from three other projects: TPAM (*Transition Planning and Assessment Model*) [2], GNOSIS (*Global Network Operations Survey and Information Sharing*) [3,4] and Aqueduct [5], a semantic MediaWiki extension. Cyber Security Research Technology System (CSRTTS) is an additional DART3 descendent, which has all the DART3 features but is tailored for use by another agency.

REQcollect combines the best components of these previous systems: a requirements elicitation and collection tool, and a Google-like matching algorithm to identify potential transitions of R&D projects that match requirements. Figure 1 shows the ancestry of REQcollect.

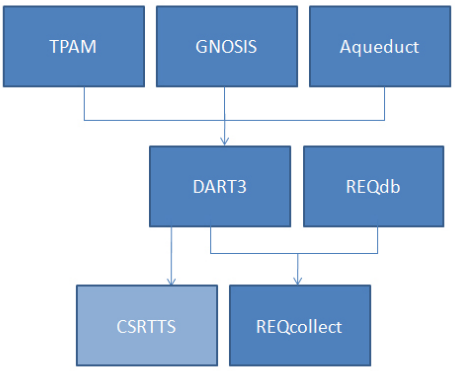

Figure 1 REQcollect Ancestry

2. ANCESTOR TOOLS

2.1 DART3

2.1.1 DART3 Summary

DART3 (*DHS Assistant for Research and development (R&D) Tracking and Technology Transition*) is a web-based tool designed to capture US federally funded research and development (R&D) projects descriptions and Department of Homeland Security (DHS) Cyber Security and Communications (CS&C) R&D requirements. DART3 automatically matches projects and requirements and assists in planning a set of transition activities to accelerate the deployment of relevant R&D results to CS&C operational systems.

Before the development of DART3, the process of matching a DHS requirement to a funded project was extremely labor intensive. Many funding agencies store their project data in a relational database, and each database may have a unique schema. For instance, in some project databases, projects are described by an "abstract" while in another database, projects may be described by a "description." Therefore, an individual researching the

978-1-4799-2504-9/14 $31.00 © 2014 IEEE
DOI 10.1109/HICSS.2014.599

relationship between projects and requirements must understand the vernacular used by various agencies. There is also much repetitive work in searching for projects across different agencies' project database interfaces, if such an interface even exists. Furthermore, the person completing the requirement/project matching must have knowledge of the existence of all such databases. This is a time and labor-intensive process and many projects may not even be discovered.

DART3 runs an automated matching algorithm based on the Apache Lucene search engine to complete a full text search over project descriptions and requirement keywords to suggest priortized lists of matches between requirements and projects. After matches are made, DART3's semantic wiki interface allows users to select elements and characteristics for technology transition. Transition activities are then generated by DART3 for tracking, funding, and scheduling during the transition process. New projects and requirements can be imported, transition activities are fully customizable and tracking output may be exported.

Because DART3 is a semantic MediaWiki-based tool, its interface is familiar to users, is easy to use and offers semantic links between data without imposing a complex data classification system. The centralization and standardization of project and requirement data provide easy access and automated, suggested matches and discovery. The extensibility allows it to be tailored to other projects and requirements outside of CS&C. Figure 2 shows DART3's main page.

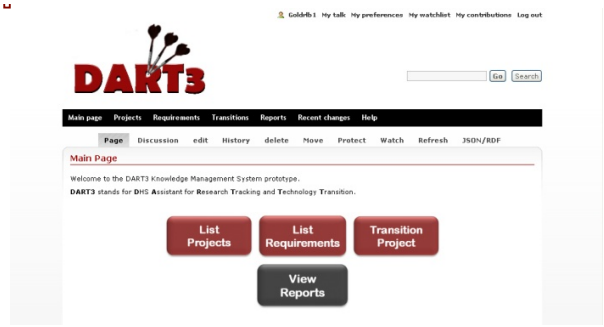

Figure 2 DART3 Main Page

2.1.2 DART3 Evolution

DART3 is a combination of GNOSIS and TPAM, two prior prototypes developed in the 2003-2008 timeframe. It also incorporates Aqueduct, a semantic MediaWiki extension.

In 2003 TPAM was originally developed under a Cross-Enterprise Technology Internal Research and Development (IRaD) project at the Johns Hopkins University Applied Physics Laboratory (JHU/APL) focusing on transition assessment and planning. It was designed to assist in the evaluation of the suitability of various aspects, or elements, of a software prototype for transition to another application development project. Effective transition requires an identification and understanding of the current state of relevant aspects of the existing software as well as the target to which it is being transitioned. The output of TPAM was the identified transition activities to be performed to successfully and smoothly execute the transition. Several elements of transition were identified including: architecture, code, design, user interface, requirements, concept and algorithm. Any given transition could involve one or more of these elements. A default set of elements, characteristics, conditions and activities for transition were supplied by TPAM but were also completely user-configurable. Figure 3 shows a TPAM screenshot.

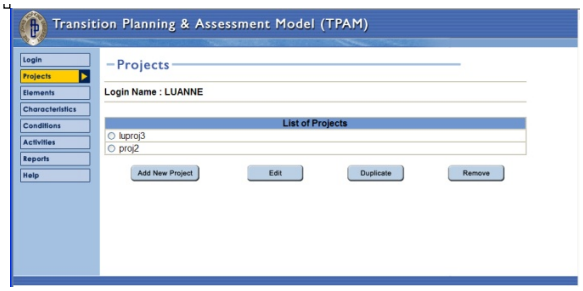

Figure 3 Transition Planning and Assessment Model (TPAM)

The development of GNOSIS began in 2008 and was funded by several agencies as a survey repository from multi-agency projects that produce cyber tradecraft. GNOSIS was a web-based, semantic wiki through which users could record details of cyber R&D projects. Data was entered manually, such as from project management summaries, conference talks, or from journal articles. The information was then available for browsing, cross-referencing and viewing via ad hoc semantic connections. GNOSIS was to be used as a single repository of research projects. At its inception and testing, National Science Foundation (NSF) projects and other unclassified R&D were recorded. Figure 4 shows the GNOSIS main page.

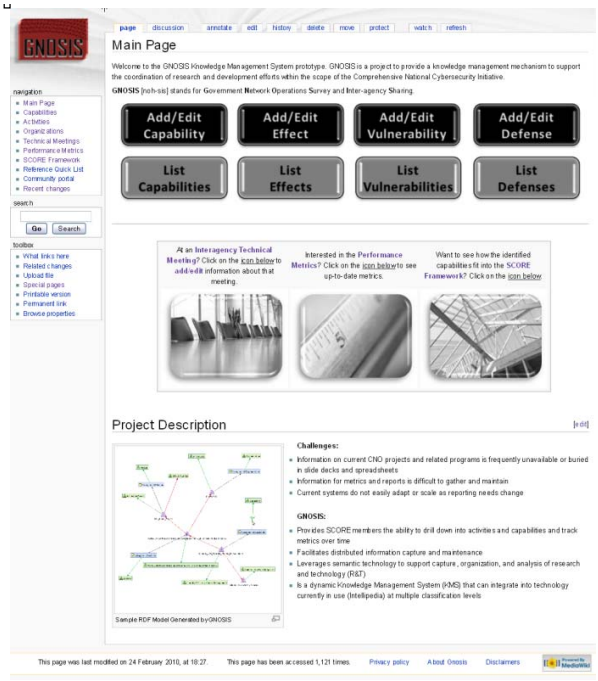

Figure 4 GNOSIS Main Page

TPAM and GNOSIS were combined in 2011 with additional features for requirements collection and requirements-projects matching. This repurposed system was named DART3 and was used internally at DHS allowing for CS&C to identify, facilitate

and track technology transition from multi-agency sponsored R&D directly into operations.

Additional features have been developed for DART3 through the use of Aqueduct. Aqueduct is a Media Wiki extension and widget framework developed through funding from several agencies since 2009. Aqueduct allows users to visualize, analyze, discuss, and enhance semantic data by viewing it collaboratively through a wiki. Aqueduct distributes queries to semantic datastores outside of the wiki and aggregates the data to display in the wiki through JavaScript widgets. Originally, GNOSIS, and thus DART3, was built upon Semantic MediaWiki [6], a separate semantic extension to MediaWiki, the wiki software that powers Wikipedia [7]. Aqueduct replaced Semantic MediaWiki in order to enhance DART3's flexibility and enrich DART3's user experience. Aqueduct provides DART3 with federated database capability and horizontal scaling, an externally queryable SPARQL interface, inline editing of data, reports and metrics, interactive query interfaces, and exportable spreadsheets.

2.1.3 DART3 MATCHING

As described in Section 2, DART3 provides the capability to automatically match DHS requirements to funded R&D projects. All project information and DHS requirements have been exported from their original databases, and imported into DART3. With all of the project and requirement data stored in DART3 in a centralized and uniform manner, a matching algorithm can be run.

The automatic matching is completed by doing a Google-like search over project titles, descriptions, and keywords, where the query is made up of the requirement keywords. The query is expanded through use of a fuzzy query and synonym expansion. A fuzzy query is a query in which the correct search is made regardless of minor misspellings of the query terms. Synonym expansion means that not only is a requirement keyword searched, but all of its synonyms, as well. Synonym expansion is particularly useful for DART3, as different users may apply different keywords to requirements or projects that have similar meaning.

The automatic matching algorithm is performed in a Java program outside of the wiki environment. A SPARQL query is performed on DART3 which returns all of the project and requirement information for input into the matching application. Apache Lucene is a text search engine library, written in Java, used to implement the matching algorithm [8]. An index of all project descriptions, titles, and keywords is created. A fuzzy query with synonym expansion is completed using the requirement keywords, and the projects with the highest scores are used to complete the *Suggested Projects to Partially Meet This Requirement* and *Suggested Requirements* sections of the requirement and project pages, respectively. The synonyms are found using WordNet [9]. The matches are then written back into DART3 using another SPARQL query. As the project and requirement information may change, the automatic matcher is run on a regular basis to update the matches. Figure 5 illustrates a DART3 Requirement page with suggested Project matches. Note that the tables which appear on the page will dynamically change based on the current matches available, because they are directly driven by the data in the underlying semantic datastores.

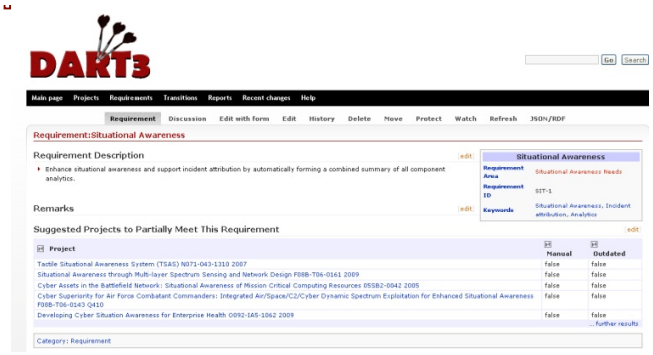

Figure 5 DART3 Requirement-Project Matches

2.2 REQdb

2.2.1 REQdb Summary

REQdb is a web-based tool designed to capture DHS Office of Cyber Security and Communications R&D mission requirements. REQdb assists in the requirements elicitation process by storing information about requirements including name, description, source, constraints, category and priority. It also keeps a running changelog for tracking the evolution of requirements. REQdb also provides the ability to archive requirements and display historical, printable reports and exportable spreadsheets.

Before the development of REQdb, requirements were recorded in flat document files. Document versioning was the only method of tracking requirement evolution and archiving. Reporting by organization or requirement area was a laborious manual process of extracting text from the documents to produce focused subdocuments. The ability to print reports organized by other fields such as priority and category was also necessary and involved cutting and pasting from document to document. Additionally, the process of matching a DHS requirement to a funded project was an extremely labor intensive and manual process.

REQdb provides a relational database repository for storing requirements and projects. Its web-based user interface allows for easy insertion and editing of requirements and projects. The reporting utilities provide an easy interface for extracting requirements and generating reports sorted on fields of the user's choosing; for example, requirements by priority, by organization, by category or by multiple fields. Requirements can be deprecated and reports can be run to include or exclude deprecated requirements. Additionally, snapshot archives can be created. REQdb runs an automated Apache Lucene matching algorithm to complete a Google-like full-text search over project descriptions and requirement keywords to suggest priortized lists of matches between requirements and projects.

The REQdb repository has a web-based interface that is familiar to users and easy to use. The centralization and standardization of requirement data provide easy access, reporting and automated, suggested matches and discovery. Figure 6 shows the REQdb main page.

Figure 6 REQdbMain Page

2.2.2 REQdb REQUIREMENTS

For the first release of both DART3 and REQdb, only DHS CS&C requirements are contained in the repository. These CS&C requirements are broken into 19 technical topic areas, roughly corresponding to the technical topic areas outlined by DHS Science and Technology (S&T). The Research and Standards Integration Program (RSI) of DHS coordinates with S&T by providing S&T with requirements for consideration in Broad Agency Announcements (BAAs); S&T, in turn, provides projects back to RSI for possible transition to operational environments.

Each technical topic area has an abbreviation and title. They include:

SWA Software Assurance
MTC Security Metrics
USE Usable Security
INS Insider Threat
NET Secure, Resilient Systems and Networks
MAL Malware Analysis
NMM Network Mapping and Measurement
IRC Incident Response Communities
CBE Cyber Economics
DPV Digital Provenance
NIC Nature-Inspired Cyber Health
INF Information Sharing
SIT Situational Awareness
SUP Supply Chain
RSP Incident Response
SCL Scalability
ADV Adversary Investigation
COM Communications
IDT Identity Management

Each requirement is then assigned to a category and assigned a number and a title, for example **SWA-1**, in the *Software Assurance* topic area:

> **SWA-1 Cloud Security** - Improve security for cloud computing where operations are heavily reliant on personal electronic devices (PEDs) and other previously non-web-enabled devices in an environment where operations and infrastructure are disconnected yet globally interconnected**.**

Figure 7 shows a REQdb Requirements Report**.**

SWA-11 Unspecified Behavior Identification
SWA-12 Guardian for Untrusted Code
SWA-13 Denial-of-Service Countermeasures
SWA-18 Embedded Attribution Techniques
SWA-19 Secure Software of Unknown Pedigree
SWA-20 System of Systems Vulnerability

Activity(Requirement ID)	Description	Owner
Adversary Investigation Needs (ADV) This section addresses the investigation of those who instigate cybersecurity incidents.		
Threat Evolution (ADV-1)	Determine the evolution and operations of cyber threat.	SWA US-CERT
Communications Needs (COM) This section deals with technical and logistic issues concerning the ability of entities within DHS to securely and quickly communicate with each other and with other entities that are charged with protecting the public.		
Encryption for Communication Devices (COM-1)	Develop an encryption algorithm for voice and video that performs at near real-time [TBD]/millisecond speed.	NCS
NGN Security for Long-Distance Communications (COM-5)	Determine the impact on the confidentiality, availability, integrity, and authenticity of priority communications when those	NCS

Figure 7 REQdb Report

3. Agile Development and Transitions

3.1 DART3 Agile Development

DART3 was developed in six agile development sprints during 2011. The red dotted box in Figure 13 shows the DART3 agile development in the context of the overall evolution of REQcollect The DART3 development objective was to promote transitions from R&D directly into operations, without the need to go through commercialization. The advantages of this approach are early use of leading edge technologies, early involvement and influence in the development, and lower costs. DART3 development brought together the advantages of TPAM (transition planning) and GNOSIS (project repository) with the addition of features including a requirements repository and matching of requirements to projects.

Each DART3 agile sprint was three weeks long with one week between sprints. The team held three scrum meetings per week to maintain schedule, and address issues and obstacles The starting feature backlog was broken out over the six sprints; as needed, features were regrouped and rearranged. The first sprint was an organizational one; the products produced were:

1. Initial feature backlog breakdown over sprints
2. OV1 (high level, Operational View) diagram (Figure 8)
3. High level process view (Figure 9)
4. Data flow diagram (Figure 10)
5. Reinstallation of TPAM environment

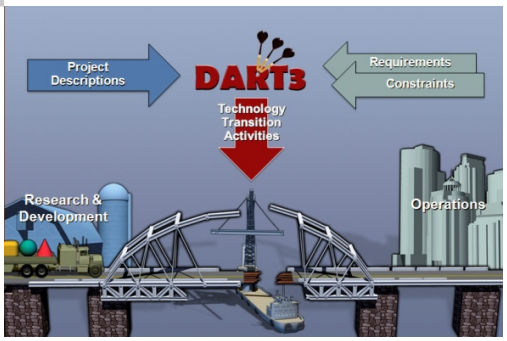

Figure 8 Operational View

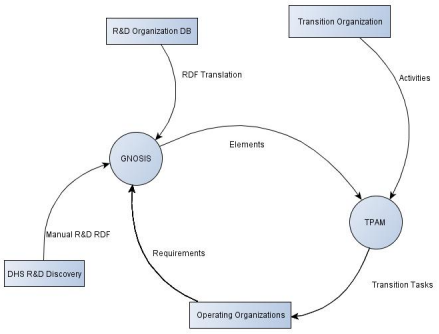

Figure 9 High-level Process View

The data flow diagram was color-coded to plan sprint goals and, after the sprint, to record progress. Figure 10 shows the planned goals for sprint 2 and Figure 11 shows the progress. In addition, a tracking burndown chart was used for each sprint, as shown in Figure 12. At the beginning of a sprint, features from the backlog to be completed this sprint were broken down further into development tasks. Each task for the sprint was assigned an owner who estimated the number of hours for the task and tracked progress as the sprint progressed. The burndown graph at the top of the chart indicated the progress towards completion for the sprint. This process allowed daily assessments of progress and an early indication if the sprint was falling behind schedule. At the end of each sprint, a working prototype of the system with the completed features from that sprint was complete, tested, and documented. Thus, each additional sprint incrementally added features towards the final system. DART3 was successfully transitioned to DHS in the fall of 2011.

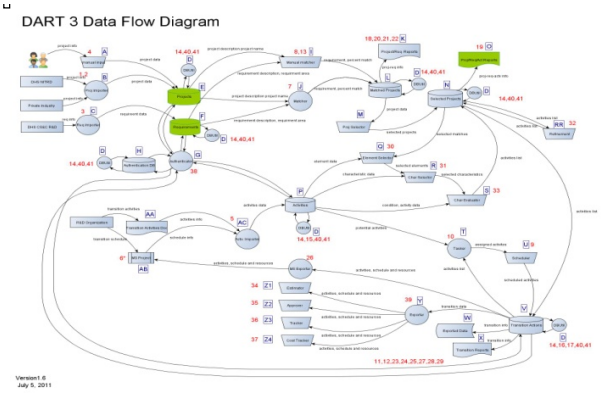

Figure 10 Sprint Goals

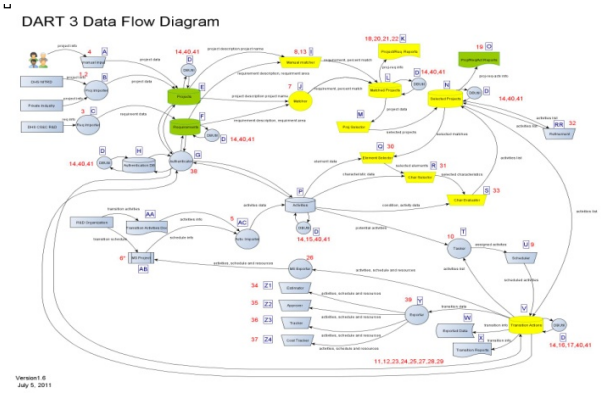

Figure 11 Sprint Progress

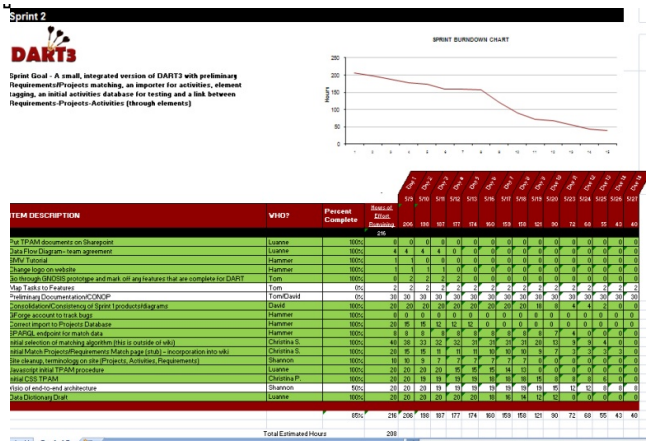

Figure 12 Sprint Burndown Chart

3.2 REQdb Agile Development

REQdb was developed in three agile sprints during 2012. Each sprint was one month long. The main objectives for REQdb were to assist in requirements elicitation and reporting. The blue dotted box in Figure 13 shows the agile development for REQdb. REQdb has a MySQL relational database that contains requirement data and a PHP web-based front end for reporting, recording and archiving. The focus of the three sprints was:

1. Database design and population
2. Reporting and export
3. User-interface and Admin functions

REQdb was delivered for transition to DHS in the fall of 2012.

4. REQcollect

REQcollect was developed in four agile sprints during 2013. Each sprint was one month long. The main objective for REQcollect was to combine the best features of both DART3 and REQdb into a single tool. The purple dotted box in Figure 13 shows the agile development for REQdb. Note the recursive progression and reuse of several tools. Typically, features were extracted from each tool during the evolution. REQcollect brings together the advantages of both DART3 (projects, requirements, matching) and REQdb (requirements

	TPAM 2003	GNOSIS 2008	DART3 2011	REQdb 2012	REQcollect 2013
Project Repository		X	X		X
Requirements Repository			X	X	X
Requirements Collection				X	X
Lucene Matcher			X		X
Transition Activities	X		X		
Semantic Wiki Interface		X	X		
MySQL/PHP Interface				X	X

elicitation tools for collection, reporting and archiving). The focus of the four sprints was:

1. Project repository and population, transition tracking
2. Matcher improvements (phrases) and parameterization
3. Advanced reporting and admin functions
4. User interface refinements and search

REQcollect was delivered for transition to DHS in the summer of 2013.

Table 1 shows the feature evolution of TPAM, GNOSIS, DART3, REQdb and REQcollect.

- DART3=TPAM + GNOSIS + Aqueduct
- REQcollect=DART3 + REQdb

Note that REQcollect combines most of the features of its predecessors with the exceptions of the transition activities and tracker and the DART3 wiki environment. A MySQL/PHP environment replaced the latter. Because each organization to which a transition is made typically has its own processes for inserting technologies, the transition tracker and activities generation functions were not included in REQcollect.

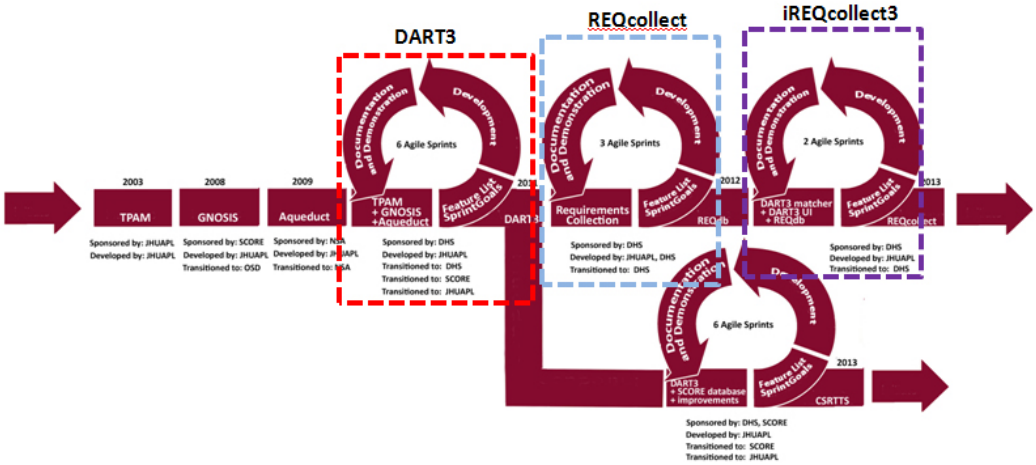

Figure 13 REQcollect Recursive, Agile Transitions

	TPAM 2003	GNOSIS 2008	DART3 2011	REQdb 2012	REQcollect 2013
Project Repository		X	X		X
Requirements Repository			X	X	X
Requirements Collection				X	X
Lucene Matcher			X		X
Transition Activities	X		X		
Semantic Wiki Interface		X	X		
MySQL/PHP Interface				X	X

Table 1 Tool Features

5. RESULTS

REQcollect has been in use at DHS since 2012 (first as REQdb). It is actively used during the collection of R&D requirements elicitation for CS&C. R&D Needs are entered, tracked and prioritized using the tool. In addition, notes and peer reviews are stored for each requirement. Various reports are generated both during the requirements gathering process and at the completion, for delivery and communication to DHS S&T. Because REQcollect also provides transition tracking, technology transitions, their status and stored documentation are all accessible from the REQcollect interface. The report archiving feature provided by REQcollect allows users to track requirements over time and over requirements elicitation cycles.

Figure 14 shows the REQcollect Requirements Listing page; from this page the user can click on individual requirements to get more detailed information. If the user is an administrator, he may also edit the requirement details. Peer reviewers may enter comments about requirements by clicking on them.

Figure 15 shows the REQcollect Projects Listing page. As with the Requirements page, depending on permissions, project details can be viewed and/or edited. Figure 16 illustrates a portion of the page that is displayed when the user views a project's details. The results of the matcher are shown in the table at the bottom of the page under the heading 'Suggested Requirement Matches.' Likewise, on an individual requirement page, suggested project matches are displayed.

Figure 17 shows the REQcollect Transitions page; it contains all the current and past technology transitions, the participating projects and fulfilled requirements, status, year, links to relevant documents and other transition details.

Figure 14 REQcollect Requirements Listing

Figure 15 REQcollect Project Listing

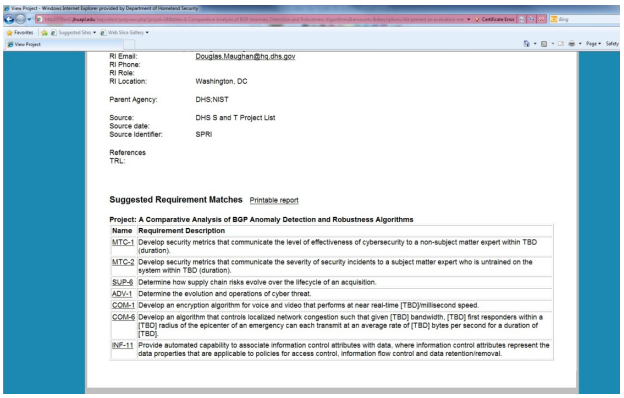

Figure 16 REQcollect Requirement Matches from a Project

Figure 17 REQcollect Transitions Listing with Tooltips UI

6. FUTURE WORK

An additional capability that would be useful for the matcher is an algorithm that uses semantic equivalence to discover matches. For instance, determining that *trustworthy computing* and *survivable systems* refer to similar ideas would result in an automatic match being made. Currently, this capability could be hard-coded if such equivalences were pre-determined. However, a more powerful tool should be considered, such as, something similar to WordNet for phrases instead of words.

Standardizing the methods used to describe R&D projects and requirements would significantly improve both the usefulness of this tool and the sharing of R&D information in general. The standardization of these descriptions would promote the use of shared database/web services that could be easily aggregated and searched. For DART3, this would provide the ability to incorporate the changing landscape of R&D projects automatically, providing updated transition matches and helping to further automate the identification of R&D that meets critical operational requirements. If R&D requirements were also standardized, then both ends of this process could be easily automated, making the system available to a much broader range of United States Government and civilian personnel. RSI will be considering approaches along these lines to assist both DHS and other departments and agencies in the identification and transition of technology that meets critical operational needs.

7. CONCLUSION

REQcollect is the tool resulting from a succession of recursive, agile development projects and technology transitions. The agile methods and subsequent transitions of each of the parent projects resulted in the evolution of more tools and then more agile development and transitions. Each tool in the progression harnessed the power and best features of the tools prior to it. As lessons were learned, the resulting tools were improved. REQcollect is a tool that was derived from five prior tools: TPAM, GNOSIS, Aqueduct, DART3, and REQdb.

8. REFERENCES

[1] Luanne Goldrich, Stephen Hamer, Christina Selby, Thomas Longstaff, "DART3: DHS Assistant for R&D Tracking and Technology Transfer," Proceedings of the 46th Annual Hawaii International Conference on System Sciences (HICSS), January 7-10 2013, Computer Society Press, 2013, pp. 5023-5028

[2] Daley, R. (2005) Transition Planning and Assessment Model (TPAM). Johns Hopkins University Applied Physics Lab IRAD.

[3] Longstaff, T., Pikas, C.K., Ferrucci, S.L., Osorno, M. (2010) Global Network Operations Survey and Interagency Sharing (GNOSIS) Demonstration. Johns Hopkins University Applied Physics Lab Report number AISD-10-29

[4] Pikas, C.K., and Ferrucci, S.L. (2010). GNOSIS Information Structure Development. Johns Hopkins University Applied Physics Lab Report number TSI-2010-003

[5] Aqueduct MediaWiki Extension. Retrieved May 29, 2013. http://code.google.com/p/aqueduct/

[6] Semantic MediaWiki. (2011). Retrieved May 29, 2013. http://semantic-mediawiki.org/wiki/Semantic_MediaWiki

[7] Wikipedia, http://wikipedia.org/wiki/Main_Page

[8] Apache Lucene-Overview. (2011). Retrieved May 29, 2013. http://lucene.apache.org/java/docs/

[9] WordNet, A lexical database for English. (2011). Retrieved May 29, 2013. http://wordnet.princeton.edu/