

PI Assist Version 3

Status And Installation Report

Marc Grushcow
NTT SYSTEMS INC.
86 Dunblaine Ave.,
North York, Ontario
CANADA
M5M 2S1
(416)575-7786

Scientific Authority: Andrea Hawton
Head/Command Effectiveness and Behaviour Support Group
(416) 635-2100 x 3036

Contract Number: W7711-007687
Call-up 7894-02

Defence R&D Canada - Toronto

Contract Report
DRDC Toronto CR 2005-104
27 January 2005

Author

Marc Grushcow

Approved by

Carol McCann
Head/Command Effectiveness & Behaviour

Approved for release by

K.M. Sutton
Chair, Document Review and Library Committee

© HER MAJESTY THE QUEEN IN RIGHT OF CANADA (2005)
as represented by the Minister of National Defense

© SA MAJESTE LA REINE EN DROIT DUE CANADA (2005)
Défense Nationale Canada

Abstract

This report describes the third development release of *PI Assist*, an experiment documentation and data archiving tool developed by *NTT SYSTEMS INC.* for the Command Effectiveness and Behaviour Section, DRDC Toronto. It describes the background and goals of the call-up, which focus on two main work items: redesigning parts of the document so that they conform with the DRDC Ethics Submission form; and, adding a feature that allows users to easily create a new *PI Assist* document from selected sections of an existing one. This deals with the common practice of having (and thus wanting to document) a series of similar experiments that are part of a larger study.

A discussion section restates the rationale behind the system – the need to protect collected data by documenting the conditions under which it was collected, and, the desire to document goals, designs, and procedures as a way of protecting the organization's investment in the intellectual property it creates. This is set in the context of other types of data archives currently envisaged and in use.

Finally, a set of recommendations are set out that describe how *PI Assist* could be used internally by the CEB Support Group itself. Benefits and limitations are identified along with the suggestion for using *PI Assist* as a vehicle for returning reports and data from research contractors to the Section.

Résumé

Le présent rapport décrit la troisième version de développement de *PI Assist*, un outil de documentation d'expériences et d'archivage de données mis au point par *NTT SYSTEMS INC.* à l'intention de la Section de l'efficacité du commandement et du comportement, RDDC Toronto. Il y est question du contexte et des objectifs de la commande subséquente, qui traite principalement de deux activités d'importance : refonte de parties du document de manière à ce qu'elles soient conformes au formulaire Demande d'évaluation de l'éthique de RDDC et ajout d'une fonction permettant aux utilisateurs de créer facilement un nouveau document *PI Assist* d'après des sections choisies dans un document existant. Le présent rapport porte sur une pratique courante visant à recourir à une série d'expériences semblables faisant partie d'une étude plus vaste (et donc à documenter celles-ci).

Une section réservée aux discussions décrit de nouveau le bien-fondé du système – la nécessité de protéger les données recueillies en documentant les conditions dans lesquelles elles ont été recueillies et la volonté de documenter les objectifs, les concepts et les procédures, de manière à protéger l'investissement de l'organisation au chapitre de la propriété intellectuelle ainsi engendrée. Cela se traduit dans le contexte d'autres types de données envisagés et utilisés actuellement.

Finalement, un ensemble de recommandations est présenté afin de décrire comment le Groupe de soutien d'ECC pourrait se servir à l'interne de *PI Assist*. Les avantages et les limites sont mentionnés, ainsi que la suggestion d'utiliser cet outil comme moyen permettant de retourner à la Section les rapports et les données d'entrepreneurs en recherche.

Executive summary

PI Assist is an experiment documentation and archiving tool developed by *NTT SYSTEMS* for the Command Effectiveness and Behaviour Section, DRDC Toronto. Its goals are to collect, document and preserve the intellectual property created in the process of conceiving, designing and conducting research studies.

This report describes new features that were added for Version 3, discusses the system in the context of other archiving systems and presents recommendations for using the program within CEB.

Sommaire

PI Assist est un outil de documentation d'expériences et d'archivage de données mis au point par *NTT SYSTEMS* à l'intention de la Section de l'efficacité du commandement et du comportement, RDDC Toronto. Ses objectifs sont de recueillir, de documenter et de protéger la propriété intellectuelle engendrée lors du processus de définition, de conception et de conduite de travaux de recherche.

Le présent rapport décrit les nouvelles fonctions de la version 3, discute du système dans le contexte d'autres systèmes d'archivage et présente des recommandations relatives à l'utilisation du programme à ECC.

This page intentionally left blank

Table of contents

Abstract.....	i
Résumé	ii
Executive summary	iii
Sommaire.....	iii
Table of contents	iii
Background and Goals.....	2
Work Done for Version 3	2
Discussion.....	3
The Context	3
The Technology	4
Experience	4
Recommendations for Using <i>PI Assist</i> at CEB.....	5

This page intentionally left blank

Background and Goals

This report deals with the creation and installation of *PI Assist* Version 3. The initial version was created as a proof of concept, demonstrating how experiment metadata could be easily collected and stored along with an archive of related experimental files including experiment support files, analytical results, design documents, etc. A second version was created and installed for selected scientists and support staff as a field trial. Users were asked to identify features that needed improvement and to suggest enhancements that they felt would make the system more user-friendly.

We reviewed the responses and developed a set of work items that would be incorporated in this, the third version. Version 3 is intended to be the last “development version” of the system. That is, it now contains all of the features thought necessary to meet its stated goals and to satisfy user requests. Any further enhancements will be solely in response to needs identified as the result of actual use.

The remainder of this document describes the work done to satisfy Call-up 7894-02; discusses *PI Assist* in CEB and more generally in the context of experiment archiving systems; and presents recommendations for using the program in CEB.

Work Done for Version 3

Version 3 was meant to address certain usability issues and to provide features that would provide value added to scientists as a way to encourage them to adopt the system. These fell into two main areas: document design and support for related experiments.

The *PI Assist* document was redesigned so that a major section now mimics the layout and content of the standard DRDC ethics submission. This was accompanied by the development of an Ethics Report that extracted ethics section data and formatted it in a way that matches the expected submission document. With these changes, Scientists could use *PI Assist* to enter a significant percentage of the background metadata and at the same time produce their ethics submission.

Documentation and the context sensitive help system were updated to reflect these changes and the sample document and other existing *PI Assist* documents were restructured to match the new document design.

The second major enhancement addressed the fact that many research projects consist of a series of related experiments in which the general experimental environment is constant but the underlying design (conditions, dependent variables, etc.) is manipulated. This implies that much of the metadata is the same for all experiments in the series. We developed a selective copy mechanism. This lets researchers create a new *PI Assist* document from an existing one. Thus, much of the tombstone data, overview information and ethics submission could be automatically transferred from the last experiment into the document for the next experiment. This could include any selected sections including lab facilities and software, data cleaning

procedures, analysis overview, and so on. Copied information is colour coded in the new document to indicate that it has been copied and not modified. Updating a section to reflect the actual characteristics of the new experiment automatically removes the colour coding. An other feature allows the user to explicitly indicate that a copied section is correct as it stands (i.e., no changes are required), and this also removes the colour coding.

Version 3, along with the revised documentation, was installed on all CEB office machines, both for scientists and support staff.

Discussion

PI Assist was created to capture both the metadata that describes an experiment and the data and analysis results that derive from it. Without it, or some other equivalent set of documentation procedures, the data ultimately becomes useless as the organization forgets the conditions under which it was collected.

Experiment designs, the procedures for mounting experiments and managing and cleaning the data, along with results and conclusions are also valuable organizational assets. They can inform new work and facilitate the efficient running of the lab itself. *PI Assist* provides a formalized mechanism for collecting this information so that it can be easily accessed and reused.

PI Assist can be used by both scientists and support staff in the normal course of their work. In doing so, they create a complete description of the experiment, from goals to analysis results and conclusions. The resulting file and archive can be mounted on a web server thus providing shared access to the description and data. Finally, the *PI Assist* document could also be formatted as a DRDC report.

The Context

There is a growing international awareness of the need to archive and share research data and metadata. It is driven by a number of concerns and goals. These include the desire (or requirement) to protect and publish data collected at public expense, and the realization that shared data represents a valuable resource in support of both public and private research.

Both of these concerns were addressed in Canada by the National Data Archive Consultation (<http://mmsd1.mms.nrcan.gc.ca/archives/>), Phase I of which completed in 2002. There are several examples of public archives including the archive of social science survey data located at the Inter-University Consortium for Political and Social Research (ICPSR) (<http://www.icpsr.umich.edu/>).

While most existing archives are self contained, others contain links to other archives and databases, thus creating a web of research data. An example is Blueprint (<http://blueprint.org>), funded by Genomics Canada and its BIND database (<http://bind.ca>) of molecular interactions (discussed further below).

The systems envisaged and implemented above are significantly larger and more sophisticated than *PI Assist*, but they address similar concerns. And these concerns are valid both for CEB and DRDC in general.

The Technology

PI Assist is based on XML, an open standard technology used across the world for representing structured documents. These include experiment descriptions, social science surveys and survey data, molecular interactions, purchase orders, credit card transactions, etc. ICPSR uses it to store surveys and transforms these XML documents on demand to create web pages. Blueprint uses XML to import and export information to and from the BIND database. Because the data is structured, XML documents can be searched in an intelligent way (e.g., documents with 'Baranski' as principal investigator, but not as an author in a reference).

PI Assist though its design and underlying technology allows CEB to collect and store metadata and data in a way that allows it to be shared, reused and searched. Its use allows us to answer questions like: in which experiments did we collect ECG data, or use the GTP task; when did we run the "Trust Study,"; or, what analysis protocol did we use on the "TITAN Advisor Study."

Of the archiving systems that we have found, *PI Assist* is unique in that it provides a metadata editing environment that allows users unfamiliar with XML to create complex documents. These can be developed by teams of scientists and support personnel over the course of designing and running a research study, and the archive and document are a by-product of the doing the work.

Experience

ICPSR and Blueprint both demonstrate the value of data and metadata archiving and XML as an underlying document representation. These benefits are consistent with CEB's desire to collect and manage this kind of information in a way that benefits the organization over the long term. The key issue is, "how to create these documents?" *PI Assist* was designed to do this in a way that can be used by all CEB staff in the normal course of their work. However, our experience has not been positive due largely to the fact that the scientific staff are not enthusiastic about changing the way that they do their work.

This leads us to a curatorial approach. That is, originators provide information, and curators edit, complete, and format it. This is ICPSR's approach. Submitters provide ICPSR with a written description of their survey projects, the survey forms, the data, and a description of data coding and format. ICPSR's curators massage this and produce the document and datasets that ultimately wind up in the archive.

Blueprint's (BIND's) approach is similar. Once a scientist submits their molecular interaction paper to a journal, they can go to the BIND web site. Here they simply enter nine pieces of information that characterizes the interaction that they are reporting. It is registered and a BIND ID is created for it. Blueprint's curators complete the BIND document. This includes

references to the publication, references to other publications that support or refute the claim, links to related information in other databases, etc. The full report on a BIND ID is frequently over 12,000 lines long (XML). It should (unfortunately) be noted that researchers rarely go to the trouble of entering the nine pieces of registration information and Blueprint usually receives it by arrangement with the journal publishers.

Recommendations for Using *PI Assist* at CEB

At this point, it may be best to recognize that the support group is the one that is ultimately concerned with the long term maintenance of experimental data and metadata. The implication is that the curatorial work falls to it. This in fact need not be burdensome, and even limited completion of the *PI Assist* document will provide significant benefits.

Using *PI Assist* for a completely new study might work like this:

- 1) Start a new document and provide document author, PI, and title information.
- 2) When the ethics document has been approved, cut and paste sections into the Overview section. This will be straight forward since there is a one to one relationship between the documents at this level. The benefit is that this easily provides a good description of the experiment and its goals.
- 3) Add information about other collaborators, as applicable.
- 4) In collaboration with the PI, complete the Method section (Subjects, Apparatus, and Design). This has to be documented in any case in the course of setting up the experiment environment.
- 5) When the subjects have been run, collect and zip the raw data and add it to the archive.
- 6) Document and perform the data cleaning procedures. Zip the clean data and add it to the archive.
- 7) Perform the standard analyses, zip the (JMP) files and add them to the archive. Pass these files back to the PI for further analyses.
- 8) Collect other support files (target files, GUI's, batch files, etc. as desired and appropriate) and add these to the archive.
- 9) If experimenters' run notes are paper, scan to PDF and add to archive. If they are Word or similar files, add them directly.

The net result would provide a good description of the experimental design, the procedures used to conduct the experiment, the raw and clean data, first level analysis results, and related files. Depending on the degree to which experiment support files were collected, it should actually be possible to replicate the experiment at a later date. Missing would be the PI's conclusions, statistical results, and any documentation that they might have written to support their analyses. Note that the document and archive would meet the organizational goals of documenting experiment design and intent and protecting the raw data so that it could be used in later retrospective studies.

This approach is sufficient, but obviously less than ideal. It could be augmented by having external research contractors use *PI Assist* as the vehicle for submitting their experiment report and data back to CEB. Since CEB could specify the level of documentation as part of the contract agreement, the returned package could be used to create a Technical Report that would go to the library. The document and archive itself would be a valuable deliverable for the Scientific Authority, who might thus be encouraged to use *PI Assist* in their own work.

The system, as delivered and installed is operational and will continue to be supported and maintained by *NTT*.

UNCLASSIFIED

DOCUMENT CONTROL DATA <small>(Security classification of the title, body of abstract and indexing annotation must be entered when the overall document is classified)</small>		
1. ORIGINATOR (The name and address of the organization preparing the document, Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's report, or tasking agency, are entered in section 8.) Publishing: DRDC Toronto Performing: NTT Systems Inc., 86 Dublaine Ave., North York, ON, M5M 2S1 Monitoring: Contracting: DRDC Toronto		2. SECURITY CLASSIFICATION <small>(Overall security classification of the document including special warning terms if applicable.)</small> UNCLASSIFIED
3. TITLE (The complete document title as indicated on the title page. Its classification is indicated by the appropriate abbreviation (S, C, R, or U) in parenthesis at the end of the title) PI Assist Version 3: Status and Installation Report (U)		
4. AUTHORS (First name, middle initial and last name. If military, show rank, e.g. Maj. John E. Doe.) Marc Grushcow		
5. DATE OF PUBLICATION <small>(Month and year of publication of document.)</small> January 2005	6a NO. OF PAGES <small>(Total containing information, including Annexes, Appendices, etc.)</small> 14	6b. NO. OF REFS <small>(Total cited in document.)</small>
7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.) Contract Report		
8. SPONSORING ACTIVITY (The names of the department project office or laboratory sponsoring the research and development – include address.) Sponsoring: Tasking:		
9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant under which the document was written. Please specify whether project or grant.) Contract No. W7711-007687	9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)	
10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document) DRDC Toronto CR 2005-104	10b. OTHER DOCUMENT NO(S). (Any other numbers under which may be assigned this document either by the originator or by the sponsor.) Call-up No. 7894-02	
11. DOCUMENT AVAILABILITY (Any limitations on the dissemination of the document, other than those imposed by security classification.) Unlimited distribution		
12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, when further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.) Unlimited announcement		

UNCLASSIFIED

UNCLASSIFIED

DOCUMENT CONTROL DATA

(Security classification of the title, body of abstract and indexing annotation must be entered when the overall document is classified)

13. **ABSTRACT** (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

(U) This report describes the third development release of PI Assist, an experiment documentation and data archiving tool developed by NTT SYSTEMS INC. for the Command Effectiveness and Behaviour Section, DRDC Toronto. It describes the background and goals of the call-up, which focus on two main work items: redesigning parts of the document so that they conform with the DRDC Ethics Submission form; and, adding a feature that allows users to easily create a new PI Assist document from selected sections of an existing one. This deals with the common practice of having (and thus wanting to document) a series of similar experiments that are part of a larger study. A discussion section restates the rationale behind the system – the need to protect collected data by documenting the conditions under which it was collected, and, the desire to document goals, designs, and procedures as a way of protecting the organization's investment in the intellectual property it creates. This is set in the context of other types of data archives currently envisaged and in use. Finally, a set of recommendations are set out that describe how PI Assist could be used internally by the CEB Support Group itself. Benefits and limitations are identified along with the suggestion for using PI Assist as a vehicle for returning reports and data from research contractors to the Section.

14. **KEYWORDS, DESCRIPTORS or IDENTIFIERS** (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

(U) experiment documentation; archiving; new features

UNCLASSIFIED