CORA T20: Scoping Study for the Development of a Historical Costing Database

ISR Report 6091-01
Version 2.0
31 March 2017

Presented to:

Dr. Binyam Solomon
Senior Defence Scientist
Defence Research and Development Canada
Centre for Operational Research and Analysis
MGen. Pearkes Bldg., 6 CBS
101 Colonel By Drive, Ottawa
K1A 0K2

Prepared by:

International Safety Research
38 Colonnade Road North
Ottawa, Ontario
Canada K2E 7J6

Contract report
DRDC-RDDC-2017-C092
March 2017

QUALITY ASSURANCE AND VERSION TRACKING

Authorization

Title	T20: Scoping Study for the Development of a Historical Costing Database	
Report number	6091-01	
Version	2.0	Signature
Prepared by	Chad Watson	
Reviewed by	Sandy Lavigne	
Approved by	Ian Becking	
Approved for Corporate Release by	Mike McCall	

Version Tracking

Ver.	Action	By	Date
1.0	Release to Client	McCall	22 Mar 17
2.0	Release to Client	Becking	29 Mar 17

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

One of the greatest challenges in determining departmental cost estimates is a lack of historical data on project schedule to build viable models and estimates to quantify schedule risk. As a measure to help correct this, Director General Major Project Delivery (Air and Land), (DGMPD (A&L)) in a joint effort between the Centre for Operational Research and Analysis (CORA) and the Centre for Costing in Defence (CCD) committed to the development of a historical costing database in order to ensure the continued improvement of departmental cost estimates.

Database scoping is a straightforward activity for a trained database administrator (DBA). In general terms, scoping a database is only a matter of determining the who, what, where, when and why (5Ws) of the data capture. This scoping study commenced with enough ambiguity in the 5Ws that knowledge generation required two to three stakeholder sessions to reduce the problem to an acceptable level of ambiguity; not all ambiguity was removed. In fact, it was a critical finding of the stakeholder group to realize that a suitable cost and schedule risk data store must allow for ambiguity to provide meaningful data over a long period such as 100 years (50 years in the past to 50 years into the future). To ensure ambiguity did not confound discussion during the knowledge generation effort, a constant push towards a data dictionary and a set of executable specifications was made. Although the database scoping study did not result in a database, the practical exercise of scoping a historical costing database resulted in a specification, later coined “Historical Estimates and Actuals Program (HEAP)” that better suited the tasks to be performed for this scoping study.

1.1 Project Scope

This project entitled “Scoping Study for the Development of a Historical Costing Database” initially implied a predominantly technical activity related to database design. Due to pervasive ambiguity in the availability, reliability and accuracy of data, the effort of this study could be more accurately characterized as a knowledge generation activity that ultimately resulted in a data model. The resulting model includes enough detail that it could be implemented as a database if a future implementer chooses that implementation.

2. METHODOLOGY

The methodology employed for the database scoping study was a multi-phased process comprised of three consultative meetings and information gathering sessions with subject matter experts (SMEs) and stakeholders from various government departments. These sessions were designed to gain the broadest possible context required for the resulting data model. Stakeholders were required to determine what data needed to be stored, along with obtaining an agreement as to the meaning and interpretation of the data elements. The meetings were intended to draw out and capture stakeholder requirements in the form of "Executable Specifications" and data requirements in the form of a "Data Dictionary" as the requirements relate to business facing requirements[1].

The concept of "Executable Specifications" was introduced to the stakeholders in session number one as a means of collaborating on specifications. What is important about executable specifications is that they require a certain level of collaborative rigour that allows them to later be automatically tested using a tool called Cucumber [1]. Executable specifications are written in a language called Gherkin [1] that is designed to be business readable, and domain specific. It allows software behaviour to be described without detailing how the behaviour is implemented.

A full list of stakeholders, SMEs, and contractors who participated in the consultative sessions can be found in Annex A.

2.1 Consultative Sessions

To facilitate an effective discussion, a set of inputs was developed and distributed to the stakeholder group prior to each of the three stakeholder sessions. The inputs included relevant sources, a succinct background, an end state, and a near term deliverable. These inputs were defined before the first stakeholder session, and then subsequently refined during each stakeholder session.

2.1.1 Relevant Sources

Several relevant references were used to generate shared knowledge and scoping study outputs. The list of identified references grew as the stakeholder group developed a shared understanding of the problem domain. The following list of relevant sources were used:

- SOW [1]
- Project Approval Directive [2]
- Internal Briefing Note [3]
- New Generation Fighter Capability: Life Cycle Cost Framework [4]
- Development of Cost Breakdown Structure for Defence Acquisition Projects [5]
- Expert Modeller on Estimating Milestone Dates [6]
- SME Opinion Data [7]

[1] Note that fully defined executable specifications include both a business facing and technology facing specifications. This work will only deal with the business facing specifications.

- Cost Risk Framework [8]

2.1.2 Background

Using relevant sources noted in Section 2.1.1, the following statements were proposed to the stakeholder group for consideration as a succinct set of background statements:

1. The department requires an accurate history of cost evolution so that it can account for impact of risk on schedule and cost in departmental financial analysis and reporting;
2. The history of cost evolution may be used for[2]:
    a. Forecasting annual cash flows;
    b. Monitoring schedule risks at regular intervals;
    c. Quantifying (monetizing) the schedule risk for inclusion into quarterly or annual reports;
    d. Building viable models to estimate schedule risk;
    e. Building viable models to quantify (monetize) schedule risk; and
    f. Studies on cost growth.
3. Costs are currently captured at the initial rough order of magnitude (ROM) estimates and at milestones prescribed by the Project Approval Directive (PAD) and Treasury Board Secretariat (TBS);
4. Gaps in existing historical data include:
    a. Historical Project Schedule Data - data for planning, budgeting and milestones has gaps;
    b. Historical Cost Estimates – data at specified milestones has gaps; and
    c. Historical Expenditure Data – gaps in price and quantity prevent translation into costing data.

2.1.3 End State

As an input to stakeholder session number one, the end state of the historical costing database was summarized as:

"A historical costing database that supports the cost risk framework developed in 2015".

2.1.4 Near Term Deliverable(s)

Due to ambiguity in the availability, reliability and accuracy of data, a conscientious *"go-no-go"* approach was adopted through the planning and development of this study. This ensured the flexibility to abandon the effort if the discussions were un-necessarily confounded by non-consensus of stakeholders. As an input to each stakeholder session, the session attendees were required to focus on the following:

- The requirement to determine the current data sources and data to be stored in the new costing database. This would involve consultation with/and agreement amongst all stakeholders as to what data is to be stored, along with an agreement as to the meaning and interpretation of the data elements; and

[2] additional uses were identified in Session 1 and are addressed in Section 3.

- A requirement to document the data sources identified and the understanding of requirements between stakeholders. The document must provide a description of the data items required, their attributes, constraints and relationships that hold between them.

2.2 Determining Database Scope

During stakeholder session number one, the attendees were asked to review the list of potential uses, add any that were missing, and then vote for the use that was most valuable. The uses were named by their paragraph number of “2d”, “2j”, and “2a”. The uses that were presented were drawn from the available references and submitted to the stakeholder group during the discussion of the background.

Once a list of usages was compiled, the session attendees were asked to elaborate on the usage by providing a name for who was currently doing cost and schedule risk analysis, what they did, and what artefacts were produced while doing so. Due to time constraints, any usage that could not be elaborated with a minimal “who, what, artefact” was removed from consideration for analysis under the contract [1].

To further focus the analysis, the attendees were asked to vote for the usage that held the most value to them in terms of a historical costing database. The votes were used to set the order (and priority) of analysis. The order of analysis resulted in 2d (Schedule Risk), 2j (Cost Estimation Risk), and 2a (Cost Risk Mitigation).

The process of identifying, elaborating and voting was used to accomplish the following:

1. Shared Understanding – The attendees varied in background. Some attendees were knowledgeable in the domain of cost and schedule risk and others were knowledgeable in the systems that hold cost and schedule risk data. The process promoted a shared understanding amongst stakeholders of the different uses of historical cost and schedule risk data;
2. Set Scope – The list of possible uses started with six items and grew to 10 during session number one. To ensure that the analysis effort was not pulled in too many directions at once, the process reduced the scope of the analysis to the three highest value usages; and
3. Set Priority – Although the item with the most votes ended up being first in the order of analysis, the group had an opportunity to influence priority to achieve goals other than highest value, such as highest risk, easiest gain or other. The group agreed to highest value as the top priority for analysis. A timeline for planning tasks was established and the results of each subsequent meeting and/or milestone can be found in Annex B.

3. KNOWLEDGE GENERATION

The stakeholder discussions were an effective mechanism to generate a consensus on requirements, reduce levels of ambiguity, and generate knowledge to ensure a shared understanding of the strengths, weakness, opportunities and threats of different data sources.

Initiating the development of a historical costing database identified current data sources that are available, reliable and accurate. Through stakeholder consultation and agreement, members collectively determined what costing data is (or should be) stored, as well as ensured alignment between members on the meaning and interpretation of data elements. By the end of the consultative process, the level of ambiguity that would affect design had been removed or reduced to an acceptable level; any remaining ambiguity was accounted for in the final design.

The process of generating knowledge resulted in an executable plan, a clear and prioritized scope, an aggressive timeline that would result in the highest value analysis, and a constant discussion of data sources.

3.1 Data Sources

Data sources was a constant topic throughout all stakeholder sessions. The most heavily discussed data sources were the Capability Investment Database (CID) and Defence Resource Management Information System (DRMIS). Others were identified in follow-on sessions and are depicted in Figure 1.

3.1.1 SWOT of CID, DRMIS, HEAP

CID and DRMIS data sources were heavily compared throughout the demonstration and stakeholder sessions. The comparison has been captured as a Strength-Weakness-Opportunity-Threat (SWOT) analysis. The following tables are not meant to be exhaustive or imperfect. They are considered the “shared understanding” of the consultation group.
To direct the analysis toward a business objective, the overall objective was provided as:

Objective: *Maintain a historical evolution of cost and schedule data to support future cost and schedule risk analysis.*

What follows is a SWOT for the CID, for DRMIS and for a HEAP.

Table 1: CID SWOT

Factor	Perspective	Comment
Internal	Strength	• CID is more complete than DRMIS for milestone dates (it has provided data to DRMIS in the past); • Contains project data going back to the mid-90s, specifically 1997; • Is perceived as 'easy'; • Does not require much training; and • Is an excellent document repository.
Internal	Weakness	• Does not have live updates; • Does not have exportable FIN data; and

		• Data quality is not guaranteed (Paucity of Data).
External	Opportunity	• C.6 digit number in the CID - comes from DRMIS. It means there is a logical link between the two systems.
External	Threat	• DRMIS is viewed as a replacement for CID (as reported by DRMIS); • ADM(Mat) is trying to bring all projects into DRMIS; • Is a VCDS product; and • Redundancy in data entry between the two systems (CID and DRMIS) means it is unclear what system has what responsibility for what data. Potential policy gap.

Table 2: DRMIS SWOT

Factor	Perspective	Comment
Internal	Strength	• Has near real-time costing data; • Compared to CID, has better costing data; • DRMIS data is highly accessible to a trained DRMIS user with appropriate permissions; • VCDS milestones in the range of 100 - 119. Mandate milestones. Defined as per the PAD; • Has a module called "Project System" that tracks project data; • Is an excellent document repository; and • Has training available.
Internal	Weakness	• Closed Projects requires the PM assigned at close to open the project to allow updates; • Re-opening capital projects is difficult; • Goes back as far as 2003 for fin info, some older migrated from FMAS/MASIS: • CID goes back to 1997 • Requires training to extract required data; and • Historical milestone dates are only as good as what it gets from CID.
External	Opportunity	• MGI 6.1 - directs PMs to manage their projects in DRMIS. (active projects); and • Older projects are not in the system • C.6 digit number in the CID - comes from DRMIS. Means there is a logical link between the two systems.
External	Threat	• Relies on CID for historical project data, CID is a VCDS product; • Requires a directive at L0 and L1 to feed the Project System Module; • Not all projects use the same milestone sequence (example: contract award, first vehicle); • Is an ADM(Mat) product; and • Redundancy in data entry between the two systems (CID and DRMIS) means it is unclear what system has what responsibility for what data. Potential policy gap.

Table 3: HEAP SWOT

Factor	Perspective	Comment
Internal	Strength	• Understands the value of data from a research perspective (whereas other systems may place more value on other activities related to project data); • Data Retention. Maintains data as long as the asset is in-service (whereas other systems archive according to their own data retention policies, e.g., 7 years); • Flexible data storage. Can maintain datasets from CID, DRMIS and SME opinion surveys. (The assumption is that everything is currently held on local storage drives); and • Targeted Data Investments and Ownership. Individual HEAP implementers can target their investments into data harvesting and publishing that most suit their needs. Once they have developed the data, they own it.
Internal	Weakness	• Collaboration. As a system for the exchange of data, a level of collaboration with other systems, departments and agencies is required. Collaboration requires time and resources.
External	Opportunity	• Data Availability. There is data available in other repositories such as CID and DRMIS; and • There is departmental support for a new tool.
External	Threat	• The move to Carling needs to reduce foot print by 70% and therefore some personally archived data may be lost; and • Personnel changes (moves or retirements). Previous work is shared by recommendation through a network of experts. As people move on, corporate knowledge is lost.

Even though only CID, DRMIS and HEAP were analyzed using a SWOT, any of the data sources could have been similarly analyzed. The SWOT was not meant to be exhaustive or imperfect, but rather an accurate capture of the “shared understanding” of the consultation group. The understanding evolved over time and improved as the consultation group began to ask increasingly more precise questions about historical data.

Although HEAP has its own SWOT analysis, the HEAP SWOT exists to draw out characteristics of historical costing data that may or may not already be inherent in other systems. Depicting HEAP separately does not imply that an implementation separate from DRMIS or CID is required; rather, it merely highlights the characteristics of an implementation that must be present in a host system, whether the host system is a new or an existing system.

3.1.2 Data Availability

At the outset of the stakeholder sessions, there was a question of data availability and it became clear that data exists in various systems. DRMIS holds historical data not only from itself, but also from the systems it was designed to replace (for example CID). Due to the number of external sources, and the relative strengths and weaknesses of each, an expert modeller needs to know each system in order to reliably extract historical data from it.

The problem of data availability was not necessarily physical availability; rather, that which was more closely aligned to dissimilarities in domain ontology's. When it comes to historical data,

different sources use slightly different terms to describe the same entities. The problem therefore, is one of knowing what question to ask the external system so that the historical data can be brought a domain focussed on historical estimates and actuals (i.e., HEAP).

3.1.3 Data Security

Data security is a question of implementation. Since HEAP does not specify an implementation, an analysis of security was not conducted. The following high level items are offered for future consideration:

- **Classified Information**: Classified data collected within a HEAP implementation is similar to classified documents created by common desktop word processing software. The restrictions and controls related to classified (or protected) documents would extend to HEAP data. A HEAP implementation, like desktop word processing software, is a "user" of an established security framework, not a provider of an established security framework; and
- **Exchanging Classified Data:** If an established security framework would allow for automated exchanges of classified data, then use of an established data marking system should be investigated such as STANAG/ADatP 4774, or Trusted Data Format (TDF) [11].

Data security is a complex topic that requires its own dedicated analysis. Due to time constraints and the reliance of data security on a physical implementation, data security for HEAP was not assessed.

4. OUTPUTS

Knowledge generation was required to scope the historical costing database. The historical costing database was meant to be defined by a set of executable specifications and a data dictionary. These are available as outputs from the scoping study. In addition to those outputs, descriptions of the Business Flow, Logical Data Concepts and Historical Estimates and Actuals Program are provided.

4.1 Business Flow

The starting point for the business flow diagram was the analysis related to "2d" (Schedule Risk) and is very "Expert Modeller" centric. Most of the Business Flow Diagram (see Figure 1) was generated from the "2d" analysis. The remaining portions (nodes 11, 12, 14, 16) were added to the diagram based on discussions with the stakeholder group or one-on-one interviews. The Business Flow Diagram is a collection of nodes (grey circles) and information flows (black arrows). Each node and flow is labelled and described (by number) in the section on Business Flow Definitions.

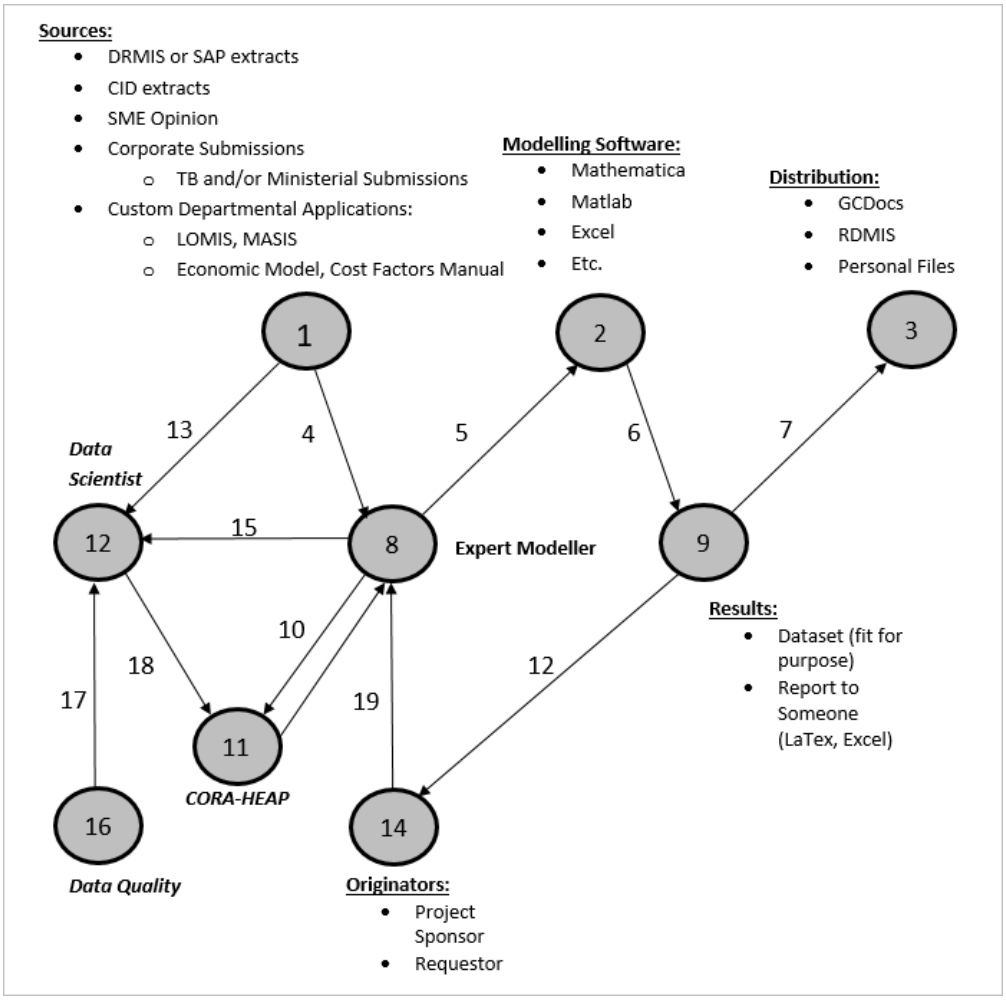

Figure 1: Business Flow Diagram

4.1.1 Business Flow Definitions

(1) External Systems. The systems that have been identified to date are DRMIS, CID, SME opinion surveys (MEOSAR), custom departmental data such as MASIS, LOMIS, Economic model, and Cost Factors Manual, Corporate Submissions such as TB and Ministerial submissions and Integrate Project Teams (IPT). As more people understand cost and risk data, more sources may be identified. Considerations include:

- If multiple sources have the same data, modellers will go to the source with the highest data quality (i.e. will consult a data source with validated data before a non-validated source); and
- Work is diverse, problems may vary, sources of data will fall in and out of favor; and
    - External systems may be proxies for source data such as Treasury Board submissions where the data may be available in more than one source.

(8) Expert Modeller. Currently the expert modeller will go to whatever source they need to get the data for their specific purpose. In all cases the scope of data is related to schedule and cost risk. The actual data elements may vary slightly from one model to another based on the specific requirement. It is expected that there will always be ***some level of transformation***. Factors that affect the transformation include: available data, available modeling tools, the training and experience of the modeller, the style of analysis, and the business objectives. The modeller is a human ETL (extraction, transformation and loading) module.

(14) Originator. The originator of the analysis is normally responsible for supplying the data. The originators may not necessarily understand the data and possible sources. As a matter of process, the originator will typically consult the Expert Modeller on the data that is required, and then request that data.

(12) Data Scientist. Analysis showed that the Expert Modeller required not only expert knowledge of the modeling software and domain of the originator, but also expert knowledge of the existing data sources. Given the variety and complexity of the identified data sources, to one degree or another, the Expert Modeller was functioning as a Data Scientist. To give appropriate prominence to the complexity of the digital data, a data scientist is reflected in the Business Flow.

(2) Modeling Software. Systems that have been identified to date are Mathematica, Matlab and Excel. The modeling software will accept source data. The modeling software:

- may approximate some gaps in source data;
- may calculate additional data; and
- produce a resulting dataset that is fit for purpose.

(9) Result Dataset. This is a dataset that is fit for purpose. The dataset will be used to generate reports that are distributed.

(3) Distribution. The reports will be sent out to various people primarily using RDIMS; GCDocs is the main means of distribution for governmental organizations. RDIMS and GCDocs allow for granular permissions. Artefacts 4-5-6 are not normally captured in RDIMS but could be.

(11) CORA-HEAP. Depicted as an example implementation of HEAP. CORA-HEAP represents an implementation of HEAP by CORA. The CORA-HEAP does not currently exist. Its presence in the diagram is to show how an expert modeller would be relieved of the requirement to go to external sources, yet at the same time, still have the flexibility using external sources if their requirements dictated it..

(16) Data Quality. Data quality is a matter of quality objectives and each external system may have different data quality objectives. Data quality in this diagram is shown to elevate the importance of the influence of data quality to HEAP data users. The data scientist would select appropriate data quality objectives for a HEAP implementation and then ensure the data going into a specific HEAP met those data quality objectives.

Artefacts:

(4) & **(13)** Outputs from source system - typically excel spreadsheets, or csv files. Output can be sent direct to the Expert Modeller or to a Data Scientist on behalf of the Expert Modeller;
(5) Input to models - typically csv or text files;
(6) Output from models - typically some form of file that can be reduced to text;
(7) & **(9)** Reports - polished copies of the results, typically in word or pdf format;
(10) HEAP inputs and outputs. Rather than going to the source system (4), the Expert Modeller may use their domain specific terminology to query an established HEAP implementation, and obtain results;
(15) Historical Data Request. An expert modeller can go directly to the external source, or go through a data scientist by submitting a historical data request;
(17) Data Quality Manual. A document outlining the quality system used to ensure data quality for a HEAP implementation;
(18) Transformed output from source system. (13) depicts outputs from source systems and (18) is that output with some level of transformation (translation) into a format that is compatible with a HEAP implementation; and
(19) Analysis Request. Expert Modellers receive analysis requests from originators.

4.1.2 Key Findings

Completion of the business flow diagram revealed the following key points:

- The Modeller will go to whatever source has the best data that fits the model requirements;
- HEAP should fit within the current application eco-system;
- HEAP data should fit an appropriate purpose; and
- HEAP will drive data requirements, relationships etc.

The specifics of HEAP will be explained later in this document.

4.2 Logical Data Concepts

Relevant sources and stakeholder discussions revealed several logical data concepts. Logical data deals with data related to business processes, categories of information, and high level obligations. It is not to be confused with physical data that is presented in Section 4.5.4.

Table 4: Logical Data Concepts

Name	Definition	Obligation	Comment
Milestone	• A serialized and consecutive point in time along a project's timeline. • Time is expressed to the granularity of a day.	Mandatory: • Textual Context Optional: • Structural Link	
Project	• A designated TB project (vote 5 or 1 as appropriate). • Consists of a Timeline (of milestones). • A project may have multiple overlapping phases. • Where ADM(Mat) is the implementer	Mandatory: • Unique ID ○ Project Metadata ○ Name ○ PCRA Level Optional: • A group of Milestones • Phases (0-N) • WBS Line Item(s)	Budget Line Items come from the MOESAR model where costs are solicited from SME opinion.
Phase	• A series of non-overlapping milestones that describe the timeline of a particular phase. • Some phase milestones may represent a project milestone.	Mandatory: • A Timeline (with a Minimum of two Milestones). • At least 1 milestone must be traceable to a project milestone.	
Planned Date	• A date value associated with a milestone that was estimated. • Time is expressed to the granularity of a day.	Mandatory: • Date Value • Date Source • Date of Evolution	
Actual Date	• A date value associated with a milestone that is the actual date that the milestone was achieved. • Time is expressed to the granularity of a day.	Mandatory: • Date Value • Date Source • Date of Evolution	
Planned Cost	• A dollar (CAD) value associated with an estimatable item	Mandatory: • Dollar Value (CAD) Optional: • Dollar Value Source • Date of Plan Evolution • Dollar Value Type • Dollar Value Confidence (%)	Dollar Value Type - MEOSAR model has EXPECTED, LOW, and HIGH dollar values Dollar Value Confidence - as a percentage
Actual Cost	• A dollar (CAD) value associated with an	Mandatory: • Dollar Value (CAD)	

Name	Definition	Obligation	Comment
	estimatable item that is considered final.	• Dollar Value Source • Date of Evolution	
Source: • Date Source • Dollar Value Source	• The name of a person, or system that is providing the milestone data. The report name, document name.	Mandatory: • Name	
Date of Plan Evolution	• A date value for the day on which the understanding of cost and schedule risk evolved. It may be coincident with the date on which a value was updated, but it is more important to capture the date on which the understanding of cost and risk changed (thereby prompting the change).	Mandatory: • Date Value	
Budget Line Item	• A project will be assigned a budget for each line item in the work breakdown structure (WBS) at project start up.	Mandatory: (one of) • Line Item Number • Line Item Name • Planned Cost(s) • Actual Cost(s)	MEOSAR project

4.3 Executable Specifications

To ensure that stakeholders understood the general concept of executable specifications, the following points were highlighted:

- Executable Specifications are a bi-product of Behaviour-Driven Development (BDD);
- BDD is a process of exploring, discovering, defining, and driving out the desired behaviour of a system;
- BDD relies on conversations, concrete examples and automated tests;
- BDD creates a shared understanding of both the problem and the proposed solution a system is designed to address;
- Concrete examples provide an opportunity to challenge assumptions;
- Cucumber is an open source tool that supports BDD;
- Cucumber uses a specification language called Gherkin;
- Gherkin has a syntax that is readable by both humans and machines. It serves two purposes – it is documentation and automated tests;
- Gherkin is a business readable and domain specific language that lets you describe software's behaviour without detailing how that behaviour is implemented; and

- Gherkin key terms include: Feature, Scenario, Given, When, Then, “As a”, “I want”, and “so that”.

4.3.1 Gherkin

The Gherkin syntax varies greatly depending on the business flow (i.e., user, Modeller, Originator, Data Scientist). As an example, some basic Gherkin syntax is as follows:

Feature: Some collection of scenarios that culminate in an identifiable feature

AS A user, ***I WANT*** a valuable piece of the overall system ***SO THAT*** I can know what I don’t know.

Scenario: Use Business Domain Language
Given an executable specification written in Gherkin
When a subject matter expert reads the specification
Then the outcome, pre-conditions and circumstance of the scenario make sense in the language of the subject matter expert

Scenario: Automate Testing
Given an executable specification written in Gherkin
When a developer automates the executable specification
Then every follow-on code change can be made with confidence that the rest of the system is still behaving as desired.

Scenario: Single Source of Truth
Given an executable specification written in Gherkin
When a technical authority validates the solution in the field
Then there is a consistent and shared understanding of the system from requirement through to fielded behaviour.

As a Modeller, I want to analyze estimatable items from 50 years ago, and until 50 years into the future So that I can conduct risk analysis on any estimatable item.

Scenario: Get Actual
Given an Evolution of Estimates
When I request the Actual for an estimatable item
Then I will get the Value, Date of Record, and Source of Record

Scenario: Get Planned Values
Given an Evolution of Estimates
When I request the Planned Values for an estimatable item
Then I will get the Estimated Value, Date or Estimate, and Source of Estimate for each Plan

Scenario: Actual is most Current
Given an Evolution of Estimates that contains an actual
When I request the Most Current value for an estimatable item
Then the actual is returned.

Scenario: An Estimate is most Current

Given an Evolution of Estimates that does not contain an actual
When I request the Most Current value for an estimatable item
Then the most recent of each type-confidence[3] pair is returned

As an Originator, I want to estimate Cost, Schedule and Qualitative Items so that I can risk mitigate my project through to completion.

I want to articulate the structure of my project so that it is easy for me to maintain my project estimates, actuals and metadata.

Scenario: Highlight Project Risk
Given available risk analysis
And Also my project's estimatable items
When I view my estimates in the context of the available risk analysis
Then areas requiring risk mitigation are highlighted

Scenario: Articulate the Structure of a Project
Given a project with items and groups of items
When I organize them into a hierarchy
Then each location in the hierarch will have a structural link.
(to augment an estimatable item's context)

Scenario: Apply Estimates
Given a project with a defined structure
When I have an estimate to report
Then I can apply that estimate to any location in the structure that makes sense

Scenario: Maintain Project Data
Given a project with:
- a defined structure
- estimatable items in the structure
- non-estimatable items in the structure when an update occurs

When the Originator is viewing project data
Then the project data is presented in an intelligently structured manner

As a Data Scientist, I want to store estimatable items in a generic way so that I can deal with ambiguity in data from 50 years in the past and until 50 years in the future.

I want the structure of a project represented in a generic way so that someone who is looking at project data 50 years in the future can look back and see:
- What items were estimatable
- What data was, or was not, captured about a project.

I want to capture generic project structures
so that gaps in data can be identified, and also,
so that Data Quality for completeness can be assessed

Scenario: Retrieve Known Estimatable Items
Given stored estimatable items

[3]The type-confidence pair is discussed in conjunction with HEAP.

When I request an estimatable item
Then both the ***Value*** its ***Textual Context*** is returned
Examples:

Table 5: Scenario Examples for Estimatable Items

Value	***Textual Context***
50$	Project ABC WBS Line Item 1.1.1
75$	Project 123 Activity 1000
100$	Project 9 CBS 6.2
125$	Project W55-2 Q2 Cash Flow
15 Jan 2002	Project DEF Implementation Milestone
Red	Project 9932 Quad Chart Risk Assessment
4	Project 8743 PCRA Level

Table 6: Scenario Examples for New Estimatable Items

Value	***Textual Context***	***Known Estimatable Item Types***
9$	Opportunity	Cost
31 Dec 2016	Completion	Schedule
A	Project Grade	Qualitative Assessment

4.4 Historical Estimates and Actuals Program (HEAP)

A Historical Estimates and Actuals Program (HEAP) is a Cost and Schedule Risk Information Exchange Data Model (IEDM). HEAP is a model that, when implemented, aims to enable the interoperability of systems and projects required to share cost and schedule risk analysis information. HEAP achieves this by specifying the minimum set of data that needs to be exchanged in departmental or international projects. Conceivably any nation, agency or community of interest that implements HEAP would be free to expand its own data dictionary to accommodate its additional information exchange requirements with the understanding that the added specifications will be valid only for the participating nation, agency or community of interest. Any addition that is deemed to be of general interest may be submitted as a change proposal within the configuration control process to be considered for inclusion in the next version of the specification. It is this process of defining and standardizing information for the purposes of interoperability that makes HEAP implementation agnostic.

HEAP IEDM is intended to represent the core of the data identified for exchange across multiple functional areas and multiple views of the requirements. Toward that end, it lays down a common approach to describing the information to be exchanged for cost and schedule risk analysis.

4.4.1 Justification

An IEDM is a trade-off between specification and planned ambiguity. The specification defines the minimum set of data required to exchange cost and schedule risk information yet the model allows for extensions and additional data definitions where it makes sense. This allows different

implementers to extend the model to fit their needs while maintaining interoperability with other systems.

A common concern that has been identified is whether HEAP will replace CID, DRMIS, or any other of the known external systems. HEAP will not replace any external systems; rather, it will allow those systems that need to exchange cost and schedule risk data to transform their data into a commonly understood format.

At the heart of HEAP is the concept of an "estimatable item" which can really be named almost anything such as a planned, project, forecasted, assigned, or budgeted (see Figure 2). The definition of these items may vary from system to system and over time. HEAP is designed to handle this type of ambiguity by allowing unstructured entry of estimatable items alongside structured project definitions. As a result, there will not be a need to enter dummy data on future projects to ensure that the data fits.

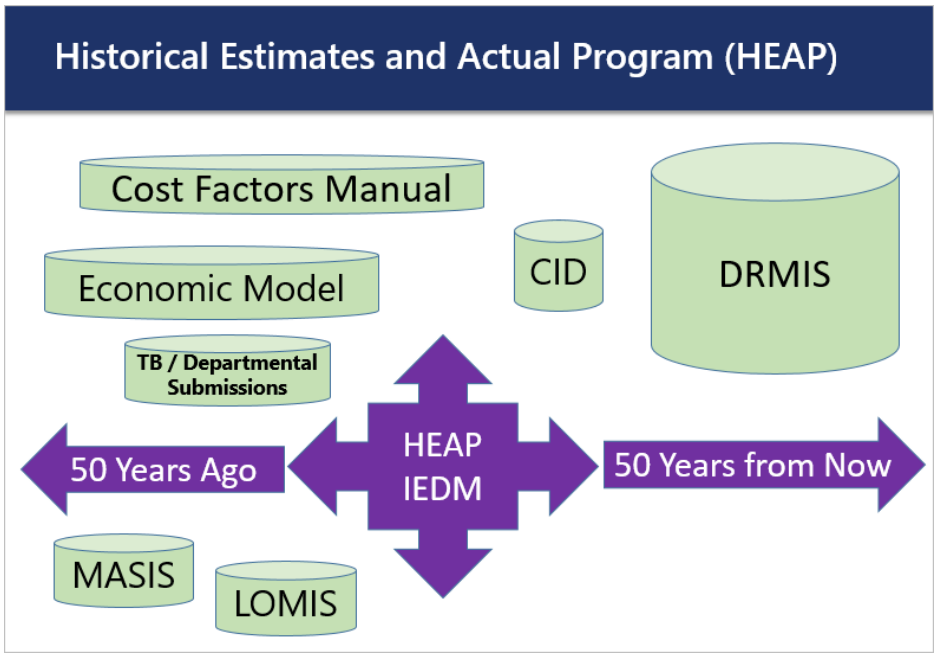

Figure 2: HEAP IEDM Across Systems and Across Time

Looking forward 50 years to the year 2067, conceptually speaking, HEAP could be implemented not only by different departments, but also by different nations who are buying the same type of equipment.

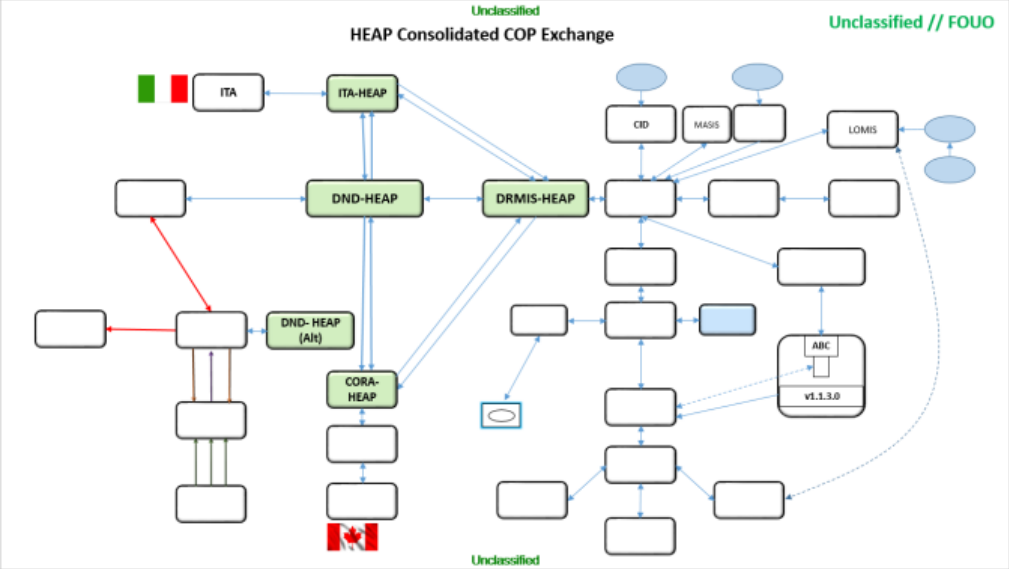

Figure 3: HEAP 2067

HEAP as an IEDM for cost and schedule risk data in the year 2067 is depicted in Figure 3 with the following notional description:

- DRMIS-HEAP – A HEAP that represents data from DRMIS that has been aggregated from a variety of legacy systems such as CID, MASIS, LOMIS etc. The DRMIS-HEAP is designed to exchange HEAP data with three other HEAPs: ITA-HEAP, DND-HEAP and CORA-HEAP;
- CORA-HEAP – A HEAP that consists of data designed for research. It is characterized not only by its data exchanges with the DND-HEAP and the DRMIS-HEAP but also by its strict data quality guidelines;
- ITA-HEAP – An Italian sponsored HEAP that is exchanging unclassified project data not only with the DND-HEAP, but also under strict data sharing MOUs with ADM(Fin) for projects where both countries are procuring the same equipment from the United States (US);
- DND-HEAP – A DND-wide HEAP used to facilitate the easy exchange of HEAP data between DND departments; and
- DND-HEAP (Alt) – An alternative HEAP used for testing and experimentation purposes.

4.4.2 HEAP Concepts

HEAP is comprised of four main concepts:

- Estimatable Items;
- Estimate Concepts;
- The concept of a Project; and
- Data Quality Concepts.

A sample project called ColPro is depicted below and may be referenced throughout the discussion of HEAP.

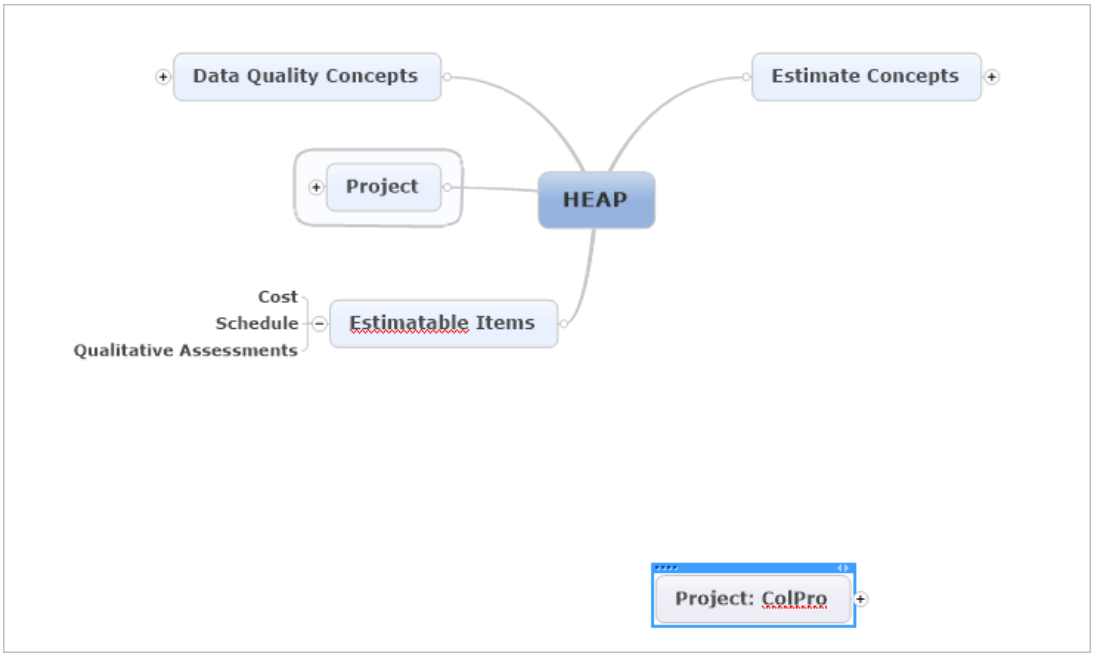

Figure 4: HEAP Main Concepts

4.4.3 Estimatable Items

At the heart of HEAP is the concept of estimatable items. Three categories of estimatable items were identified:

- Cost;
- Schedule; and
- Qualitative Assessments.

Each of these will be discussed in further detail below.

4.4.4 Estimate Concepts

Estimate concepts is defined by estimatable items such as Cost, Schedule, and Qualitative Assessments along with the concept of an Evolution of Estimates that applies to each estimatable item.

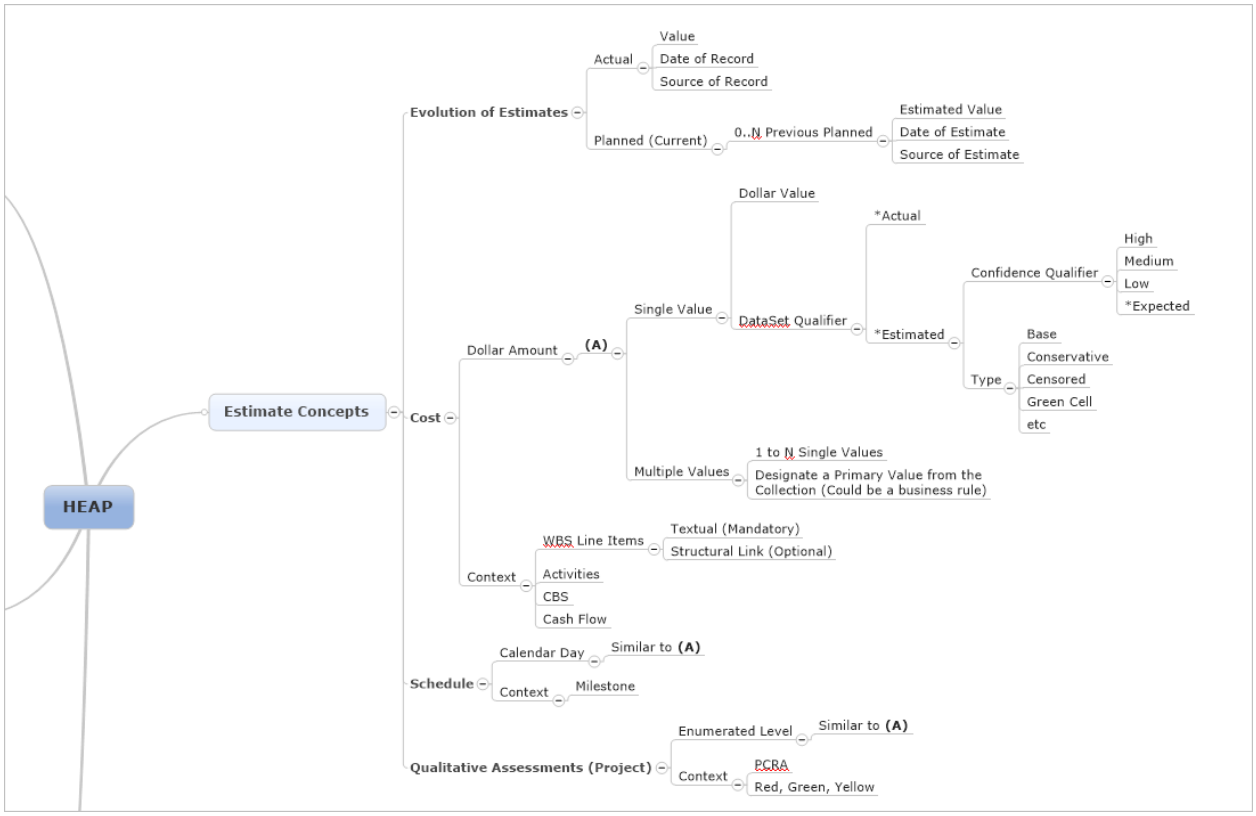

Figure 5: HEAP Estimate Concepts

4.4.4.1 *Evolution of Estimates*

An evolution of estimates is a series of planned values that ultimately culminate in an actual value. For every actual, the value, date of record, and source should exist. The planned values should have a readily distinguishable current planned value. Zero to many previously planned values may exist. Like an actual value, a planned value should have an estimated value, the date of estimate and the source of estimate.

The concept of an evolution of estimates is applicable to any estimatable item.

4.4.4.2 *Patterns in Estimatable Data - (A)*

In Figure 5 there is an (A) notation on the diagram. The (A) depicts a repeatable pattern, shown for Cost estimates, but is equally applicable to Schedule and Qualitative Assessment estimates.

The repeatable pattern says that any estimatable item (in this case, cost) may be described by either a single value, or multiple values. The implication is that implementing systems should expect to have a business rule to intelligently return a single value from a collection of multiple values, when no specific value is requested. This further implies that HEAP implementers must have a strong concept of “default” values inherent in their own implementation.

The repeatable pattern continues to define a single value as both a dollar value and a dataset qualifier.

The dataset qualifier is further defined to be either an estimated or an actual value. Note this is the point in the definition of a cost estimate where the concept for an Evolution of Estimates is seen. Whereas the evolution of estimates defined planned and actual, the cost estimate defines estimated and actual. In this context, they are semantically the same.

The repeatable pattern continues to define an estimated dataset qualifier as consisting of a type and optionally a confidence qualifier.

The type of estimate is an area where the IEDM designs for ambiguity. In some data models, each type must be explicitly defined in attributes and relationships, but in HEAP, the name of the type is sufficient. Examples of type include Base, Conservative, Censored, Green cell etc. In and of themselves, the names are meaningless to the casual reader. To HEAP, all that is required is the name, and for interoperability, a definition for the name in the local implementation's data dictionary. This should be noted as a key point. For example, if a specific HEAP is not interested in "Censored Data", there is no need to exchange that data, or have that data's definition in their data dictionary.

The confidence qualifier appeared in the context in SME opinion data. It suggests that not all estimates are the same, and at the time an estimate is created, the person or system that is providing the estimate may have a measure of confidence to assign to the estimate. Examples of confidence qualifiers include: high, medium, low, expected. Much like estimate types, confidence qualifiers are designed for ambiguity. To HEAP, all that is required is the confidence qualifier's name, and for interoperability, a definition for the name in the local implementation's data dictionary.

4.4.4.3 *Patterns in Estimatable Data – Value and Context*

Another repeatable pattern that is consistent across the estimatable items (cost, schedule, and qualitative assessments) is that they all consist of a value paired with a context. Each of these is discussed in further detail within their own sections.

4.4.4.4 *Estimatable Cost Data*

In the simplest sense, estimatable cost data for a project includes a Dollar Amount and a Context. For example, if a person is told that WBS Line Item 1.1.1 on project ColPro is $50, that is generally enough information to understand the estimate. In the example, the WBS Line Item is the context, $50 is the Dollar Amount and the project is implied. It can be said that the minimum data elements for a cost estimate include the project, the dollar amount, and the context.

The context of the cost estimate is an area where the IEDM designs for ambiguity. HEAP currently understands that cost estimates may be applied to WBS Line Items, Activities, Cost Breakdown Structure (CBS) Items, and Cash Flow items. During the scoping study there was some discussion as to whether a cost estimate could, or should, be applied to a milestone. Much like estimate types, contexts are designed for ambiguity. To HEAP, all that is required is the context's name, and for interoperability, a definition for the name in the local implementation's data dictionary.

4.4.4.5 *Structural Link Context Data*

Figure 5 depicts that any context is described in terms of a textual description (which is

mandatory), and optional Structural Link. The Structural Link is a means of providing context where the structure of a project is known. For example, if project ColPro has been defined as a hierarchy of WBS Line Items that are three levels deep, the structural link allows an estimate to be associated at a specific level in the hierarchy, even down to the N^{th} level. This has implications in how the data is interpreted in terms of aggregating data into higher levels (i.e., rolling up data).

Structural link context data is an area where the IEDM designs for ambiguity. HEAP requires textual context data, but structural link context data is optional. This means that where structure is known, it can be articulated. Where structure is unknown, it need not be articulated. Where structure standards change, or are customized, these specialized structures can be articulated. As a practical example, if one project has a milestone for "First Vehicle" and another project does not, then the project with the "First Vehicle" milestone can be articulated in HEAP, and estimates captured against that milestone.

The co-existence of optional structural context data alongside textual context data is a key enabler to HEAP being able to handle data from projects well in the past (say 50 years) to projects that are not yet known in the future (perhaps 50 years into the future).

4.4.4.6 *Estimatable Schedule Data*

Estimatable Schedule Data is almost identical to estimatable cost data except that its value and context are slightly different.

- Value – estimatable schedule data is expressed in calendar days; and
- Context – estimatable schedule data has the same style of context in that type can be defined as required. The only estimatable schedule data type that HEAP currently understands is milestones.

4.4.4.7 *Estimatable Qualitative Assessment Data*

Estimatable Qualitative Assessment Data is almost identical to estimatable cost data except that its value and context are slightly different.

- Value – estimatable qualitative assessment data is expressed as enumerations. Examples include: Levels 1 – 4, or, Stoplight colors; and
- Context – estimatable qualitative assessment data has the same style of context in that types can be defined as required. The only estimatable qualitative assessment data types that HEAP currently understands are Project Complexity and Risk Assessment (PCRA) and Red-Yellow-Green quad charts.

4.4.5 Generic Project Articulation

To HEAP, a project is a hierarchy of items and groups of items. The groups can be ordered or un-ordered. What matters to HEAP is that a project can be articulated and captured using the terms that apply to that project, and, at that time in history. Project structures change over time. So accordingly, a program like HEAP must not fall apart when a new concept like CBS appears for a project.

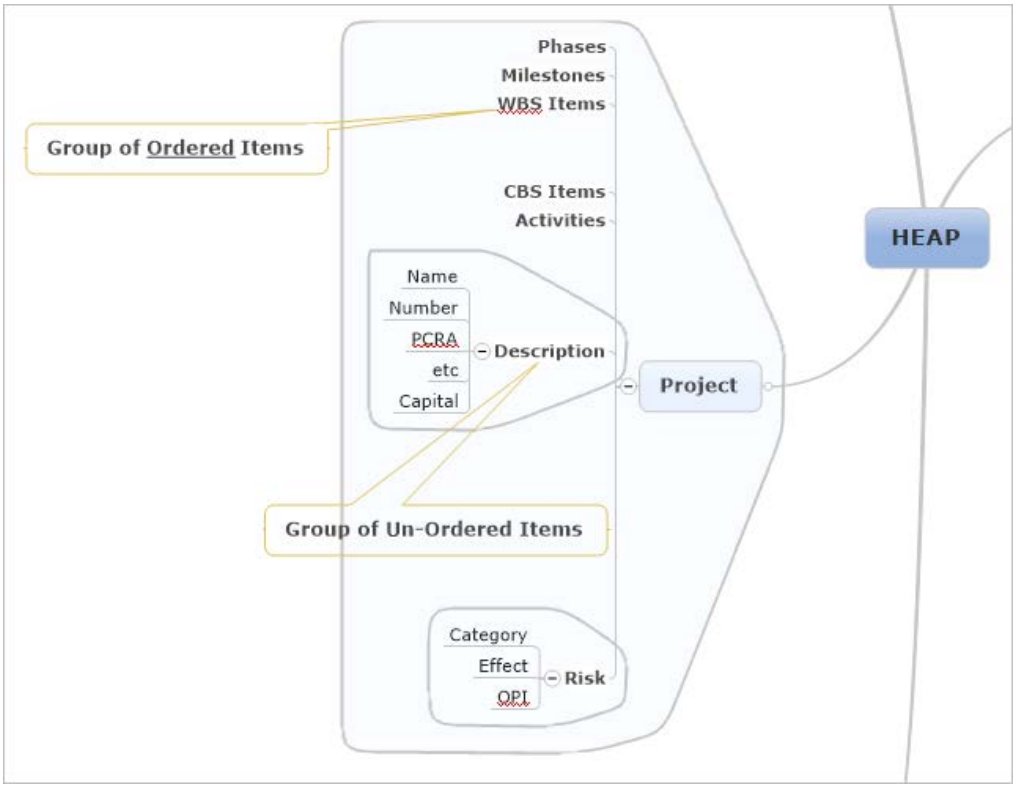

Figure 6: Generic Project Definition Capability

HEAP includes the concept that projects, both past and present can be articulated in terms of generic items and groups of items; there the groups can be ordered or un-ordered.

4.4.6 Project ColPro Example

Project ColPro is an example of HEAP's generic project articulation. In Figure 7, the Originator of the project has articulated that the project had three milestones, three WBS line items, and one descriptive item that is called PCRA.

Since a project can be generically articulated in terms of items and groups of items, the following can be said:

- Milestones is a group of ordered items (milestones), in the order of A, B, and C;
- WBS is a group of ordered items (Line Items), in the order of 1, 2, and 3;
- Description is a group of un-ordered items. Only one un-ordered item exists (PCRA);
- Milestones contain estimatable schedule data;
- WBS Line Items contain estimatable cost data; and
- PCRA is an estimatable qualitative assessment data item.

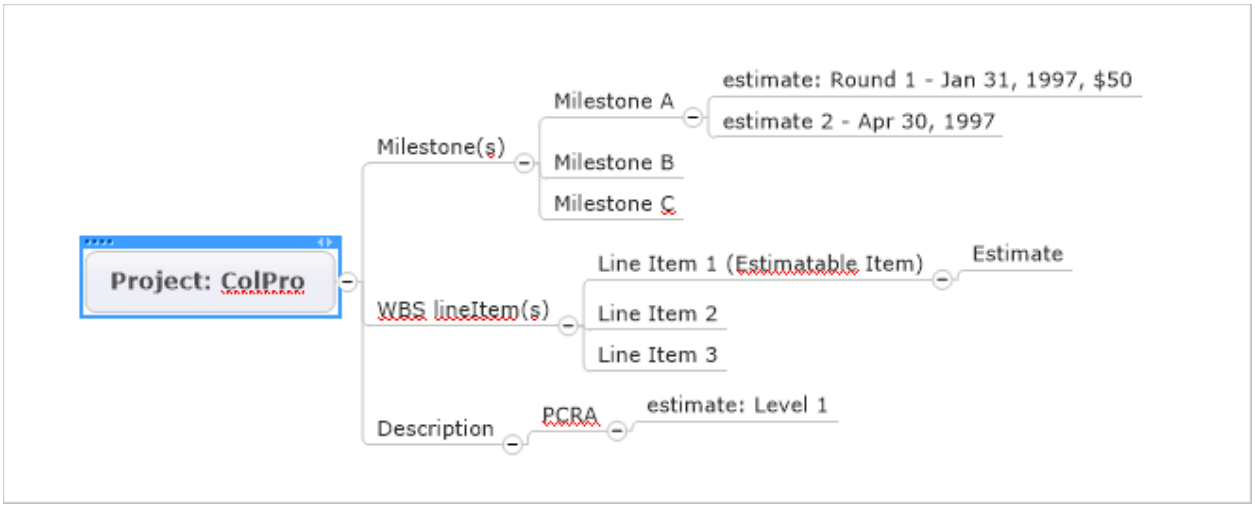

Figure 7: Project ColPro Example

One of the advantages of the ability to articulate a structure, is that the structure is able to inform a user of where there is missing data.

4.4.7 Data Quality

HEAP is designed to incorporate data quality concepts. A list of sample quantitative and qualitative data quality measures is presented below.

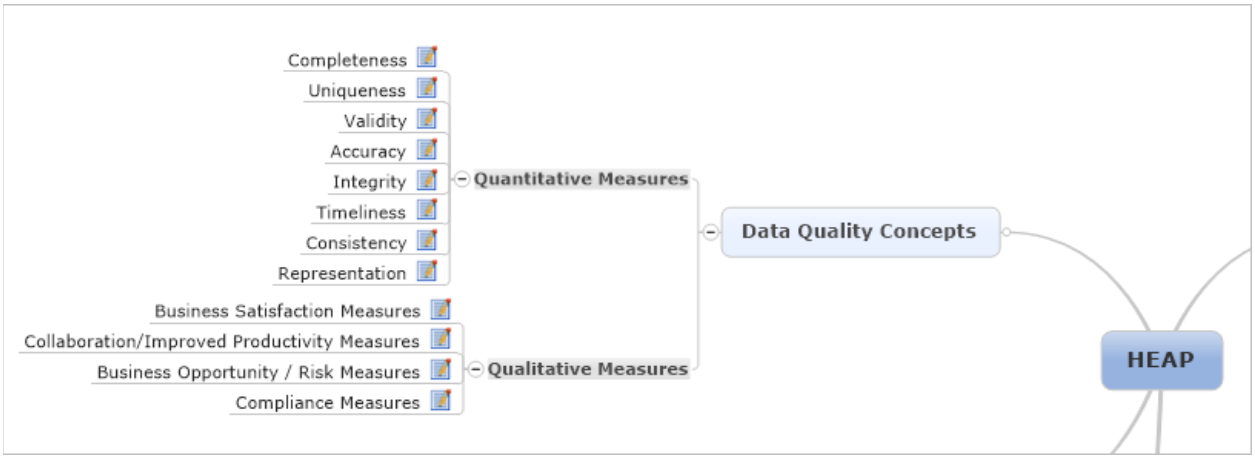

Figure 8: Data Quality

- Completeness - The degree to which all required occurrences of data are populated;
- Uniqueness - The extent to which all distinct values of a data element appear only once;
- Validity - The measure of how a data value conforms to its domain value set (i.e., a set of allowable values or range of values);
- Accuracy - The degree of conformity of a data element or a data set to an authoritative source that is deemed to be correct or the degree the data correctly represents the truth about a real-world object;
- Integrity - The degree of conformity to defined data relationship rules (e.g.,

primary/foreign key referential integrity);
- Timeliness - The degree to which data is available when it is required;
- Consistency - The degree to which a unique piece of data holds the same value across multiple data sets;
- Representation - The characteristic of Data Quality that addresses the format, pattern, legibility, and usefulness of data for its intended use;
- Business Satisfaction Measures - The increase/decrease in business satisfaction based on surveys;
- Collaboration\Improved Productivity Measures - Percent of times that data governance detects and eliminates redundant intra- or inter-departmental projects/initiative;
- Business Opportunity \ Risk Measures - Business benefit gained due to quality data or business risk realized due to questionable data. Increase in competitive analytics due to data availability and Data Quality improvements; and
- Compliance Measures - Users with access to update/influence the master data are restricted to only those employees who have need and have been approved as part of their job functions.

4.5 Data Dictionary

The role of a data dictionary is to express the logical data in more concrete terms, almost to the point of a physical database design. As indicated at the beginning of the report, an implementation was not the end state; rather a data model better suited the tasks required for this scoping study. Since the stakeholder discussions were driven by obtaining executable specifications and a data dictionary, the current version of the data dictionary is presented below.

The data dictionary consists of a glossary containing new terms and data definitions, a system description, and four domain descriptions.

4.5.1 Dictionary Glossary

Table 7 contains new terms and definitions that are referenced within the data dictionary.

Table 7: Data Dictionary Glossary

Term/Acronym	Definition
Cash Flows	A group of ordered Dollar Amounts that are estimatable. Example: 1997-Q1 Cash Flow, $1000
Close	Project Close out
DGMPD	DG Major Projects Delivery
FOC	Full Operational Capability
IOC	Initial Operational Capability
Milestones	Specific set laid out in PAD
PA	Project Approval
PA(Def)	PA Definition
PA(Imp)	PA Implementation
PAD	Project Approval Directive
PCRA	Project Complexity and Risk Assessment
Plan Versions	A term found in DRMIS. Is equivalent to an estimate

	type.
ROM	Rough Order of Magnitude
SCD	Strategic Context Document
SS(ID)	Identification – old name for SCD
SS(PPA)	Preliminary Project Analysis – old name for PA(Def)
SS(EPA)	Effective Project Analysis – old name for PA(Imp)
TBS	Treasury Board Secretariat

4.5.2 System Description

A central theme to the system description is that the proposed system is not predominantly relational. The system is a data store that includes relational database design to capture structure, and where structure is not advantageous, a navigational database design to point to the existence or non-existence of data.

It should be understood in the system's design, that the data to be stored already exists but may not be in an easily accessible format. Different consumers of the data require deep and/or wide data sets. Deep being detailed information on a few projects, and wide being basic information from a large set of projects. Currently such data sets can be difficult to collect and once collected are not readily accessible for other uses.

In order to understand the follow on conceptual groupings and data domains, a draft list of requirements and components are listed below.

System Requirements:

- Allow datasets as they are created to be stored;
- Data storage must be as frictionless as possible to make storing new datasets the default choice;
- As the number of data sets grows it must facilitate the creation of new data sets from the stored information;
- The system must allow the user to define the structure and linkages in these sets; and
- Original source of data as well as how it was inputted into the system must be tracked and the user must be able to exclude data in extracts based on this information.

Proposed Components:

- Import Component consisting of:
  - Imports excel spreadsheets; and
  - Requires entry of source context fields.
- Data Entry Component:
  - Created when data entry is the chosen input method for a dataset;
  - Can be simple or as complex as the business case for that dataset warrants; and
  - From auto-generated form to involved custom one.
- Data Extraction Component:
  - Advanced query tool;
  - Users chooses domain(s);
  - Links to other domains are not predetermined or chosen by user based on

linkable data;
    - Can choose fields, field order, filter by value, sort, group by; and
    - Export result to excel.
- Data Store Component:
    - Is divided into data domains;
    - Domains are independent of each other; links are in the source data not enforced by the data model;
    - Linked fields are fields that are common between domains;
    - Can be required on import/data entry but are not checked to be valid;
    - Each record consists of a data element and fields that define its context;
    - These must be sufficient to completely reconstruct where it came from;
    - The Structure domain can be used to reconstruct the original data field relationship structure;
    - Like list of standard PAD milestones; order of columns in a cash flow report; and
    - Facilitates the addition of fields in domains as well as additional domains.

Any additional fields must be well thought out to ensure that they are easy to use, and commonly understood. Adding additional fields in this manner will result in a more valuable data store.

4.5.3 Contextual Groupings

Each of the domains listed below include the "contextual groupings". These groupings are provided to assist the reader in logically grouping the data elements within the data domain. Descriptions of the contextual groupings are provided starting with the most re-used contextual groupings and ending with the contextual groupings that a more specific to a single data domain.

4.5.3.1 *Hierarchical Assignment Context*

The hierarchical assignment context is where structured data and un-structured data meet. The hierarchical assignment context is optional for estimatable items. Where an estimatable item's (cost, schedule, or qualitative assessment) location within a project's hierarchy is known, it can be expressed as a structural link. The structural link is optional because not all estimatable items need to be understood within a hierarchy. Furthermore, the location of an estimatable item within a hierarchy may not be known when the estimate data is acquired (a gap in data). Making the structural link optional allows an estimatable item to be captured within HEAP with structural information when it is known, and without structural information when it is not known.

4.5.3.2 *Identification Context*

The identification context allows items within a data domain to be identified uniquely.

4.5.3.3 *Source Context*

The source context captures where the data originated from. It exists to promote confidence in the data.

4.5.3.4 *Extension Context*

The extension context allows generic data domains to be augmented with specific data when it is known. HEAP implementers can decide what extensions they want to create. As extensions

appear in different implementations, and are documented in data dictionaries, the central HEAP authority can decided whether some of the extended data should be incorporated into the generic data domains. Extensions allow the HEAP IEDM to be improved over time. Extensions allow HEAP implementers to customize HEAP to their use while maintaining inter-operability with other HEAPs.

4.5.3.5 *Evolution Context*

The evolution context acknowledges that the data the data element descriptions will evolve over time; that is, there will be a conscientious progression towards an actual value. This concept of evolving data is key to HEAP. The evolution context captures enough information to be able to distinguish one evolution of data from another. In some systems, the HEAP evolution of the data may be described as a “plan”, a “version” or other term. It should be noted that although a HEAP evolution of a data element may look like a “version” or a “plan”, it is only an evolution if it is progressing toward an actual. Since HEAP is designed for risk analysis, storage of alternative plans or versions should be avoided otherwise the HEAP may become inundated with extraneous data. Evolution data, that is data that progresses toward an actual value, should be stored in HEAP. At this time, it is not the intent of HEAP to store data that will not eventually progress toward an actual or final value.

4.5.3.6 *Estimatable Context*

The estimatable context is related to the structures domain. The estimatable context allows HEAP to capture whether a location in a specific hierarchy may have associated estimate data. For example, in some projects, milestones may have estimatable costs associated with them, in other projects, milestones may not have estimable costs associated with them. The estimatable context allows HEAP to capture what portions of a project may be estimated by turning on and off flags for the different types of estimatable items.

HEAP will rely on the estimatable context to provide meaningful data quality analysis. For example, if a HEAP project specifies that milestones are estimatable for cost, yet no cost data exists within HEAP for that project’s milestones, then a gap in the data has been identified by cross referencing the estimatable context with data in the cost, schedule and qualitative assessment domains.

4.5.3.7 *Quality Context*

The quality context is a means of attributing quality to sets of data within HEAP. Should several HEAP implementations exist in the future, what will distinguish them from each other is there data quality. The quality context is a means of exchanging data quality information with the historical estimates and actuals data.

4.5.4 Data Domains

Four representative data domains are provided: cost, structures, project, and data quality. Note that schedule and qualitative assessment domains are not listed because their construction is nearly the same as cost. Not all descriptions were provided. If the name of the data element was unambiguous to the stakeholder group, then the description was not provided.

4.5.4.1 *Cost Domain*

The cost domain is more easily understood in the context of an example. Assume that Project ColPro has three milestones (A, B, C) and that the evolution of estimates will progress in three rounds.

Table 8: Cost Domain Example

Description	Milestone A	Milestone B	Milestone C
Round 1 - Jan-97	50$	$100	$200
Round 2 - Apr-97	$65 (actual)	$125	$100
Round 3 - Dec-97			

The intersection of Milestone A and Round 1 – Jan 97 could be described as:

"50$ is an estimate for Milestone 1, estimated at round one which was done in Jan of 1997 for project 123"

The cost domain of HEAP would capture the cost estimate in terms of four groupings, the evolution context, the project context, the source context, and the extension context .

Table 9: Example of Cost Domain of HEAP

Domain	Contextual Grouping	Data Elements (Element, Example, Description)
Cost	Hierarchical Assignment Context	• Structural Link (Optional) - ABC123 - Where a structure is defined, the structural link connects the cost to the node in the structure. •
	Identification Context	• Context Textual (Mandatory) - Milestone A - A descriptive name for the Cost • Project ID (Mandatory) - PID123

	Evolution Context	• Evolution (planned/actual) - planned - could also be estimated, projected, budgeted…
		• Date of Estimate:
		  ○ At Report Period - Null
		  ○ At Date - Oct-96
		  ○ For Report Period - Round 1
		  ○ For Date - Jan-97
		• Source of Estimate: Departmental Submission #345-1001
		• Dollar Amount - 50$

	Source Context	• Import flag - Y/N • Import Source • Import Date • Data Entry flag - Y/N - Yes • Data Entry User - Barry • Data Entry Date - 1 March 2017
	Extension Context (Risk, TBD)	• Risk ○ Acquisition Phase

		○ Category ○ Risk Title and Description ○ Cause ○ Effect ○ OPI ○ Probability ○ Impact ○ DRL ○ Mitigation Strategy ○ Schedule ○ Scope

4.5.4.2 *Structures Domain*

Structures - chain of item-group-item-group-etc. Used to capture project structures, but also used to capture generic structured elements such as PAD milestone hierarchy (All structures). Using the same project example, Table 10 outlines the Structure Domain.

Table 10: Example of Structure Domain

Domain	Contextual Grouping	Data Elements (Element, Example, Description)
Structures	Hierarchical Assignment Context	• Structure Link ID (Optional) - ABC123 • Project ID - 2651 - A project can have zero, one or more structures
	Identification Context	• Name (Item) - Milestone A • Type - Item (vice Group) • Structure Template (Optional) - ColPro Type Projects. Need to know what this structural row was templated from. • Version One of the ColPro Structure. Where there are more than one structure grouping in a project, the structure grouping needs a name. ○ Groups rows within the table ○ Structure name or structure link ID, one is mandatory ○ Related to domains by use of Structure Link ID
	Hierarchy Context	• Parent (Group) - Milestones • Item Order - 1
	Estimatable Context	• Cost Estimable - Yes • Schedule Estimable - Yes • Qualitative Estimable - No

4.5.4.3 *Project Domain*

Table 11 illustrates the Project Domain using the example noted in Table 8.

Table 11: Example of Project Domain

Domain	Contextual Grouping	Data Elements (Element, Example, Description)
Project	Identification Context	Project ID - 2651 HEAP Project Name
	Source Context	• Import flag - Y/N • Import Source – CID/DRIMIS/ future other • Import Date • Data Entry flag – Y/N • Date Entry User • Data Entry Date
	Extension Context (CID, TBD) CID	• Name • Number • PCRA • Category • etc • Level One Sponsor • Project Type

4.5.4.4 *Data Quality Domain*

Table 12: Example Data Quality Domain

Domain	Contextual Grouping	Data Elements (Element, Example, Description)
Data Quality	Hierarchical Assignment Context	Structural Link - ABC123
	Quality Context	Quality Measure - Completeness Assessment - 100% Details
	Source Context	• Data Entry flag – Y/N • Date Entry User • Data Entry Date

5. SUMMARY AND RECOMMENDATIONS

While the key deliverables of the historical costing database study were the data dictionary and data model, key findings were derived from the stakeholders meetings while others were derived from the outcomes of the data model. These findings include:

1. Historical data exists in DRIMIS and its previous iterations. Placing HEAP in between modellers and current data sources allows modellers to query historical data in the language of their own domain, without needing to understand the domains of the external systems. The data model structures the questions (queries) so that one can extract relevant data;
2. The data model is designed so that it can be implemented within DRMIS, or on its own: it is agnostic to current IM systems or ones that may exist in the future;
3. Different analysts have different requirements for historical data. HEAP is a data model that has enough flexibility to extract meaningful historical (estimate and actuals) data from legacy systems as well as future systems; and
4. A modeller will go to whatever source of data that best fits their needs.

As was discovered over the course of the study, the pursuit of historical data is not so much a question of whether the historical data exists; it does. With so many sources of data, the question lies in how an expert modeller (analyst) structures their data questions to extract relevant information from available sources. Currently, an expert modeller must know and understand several different data sources to be able to extract meaningful data. The outcome of the study supports an approach where an information exchange data model (IEDM) is used to provide a consistent and flexible interface to historical estimation data for analysts.

References

[1] Solomon, B. and Bouayed, Z., 2017, DRDC CORA, Contract W7714-156105/001/SV, TA-20 Scoping Study for the Development of a Historical Costing Database.

[2] Project Approval Directive. http://intranet.mil.ca/en/deptl-mgmt/project-pad.page, project-pad.pdf, project-pad_PGM113inclatP216.pdf

[3] Solomon, B. and Hu, C. 2016, Briefing Note for the Director General Major Project Delivery (Air & Land) and Centre for Costing in Defence. BN-datareq (2) (7).doc

[4] KPMG. 2012. Next Generation Fighter Capability: Life Cycle Cost Framework. NGFC LCC freamework.pdf

[5] Sokri, A. Ghergari, V. and Wang, L. 2016, Development of Cost Breakdown Structure for Defence Acquisition. Scientific report number: DRDC-RDDc-2016-R086. DRDC-RDDc-2016-R086 - DOCUMENT.pdf

[6] Desmier, P. 2017, Estimating Milestone Dates for the Delivery of Major Crown Projects. Interim Task Briefing, 13 Jan 2017, DGMOR – DRDC – CORA, Predicting_Project_Milestones.pdf

[7] Iburg, P. and Maybury, D. 2014, Scientific Letter: MEOSAR Cost Escalation Risk. DRDC-RDDC-2014-L14-0224-1455(E)

[8] Ghanmi, A., Rempel, M.,Sokri, A., Solomon, B., Ghergari, V., 2014, Cost Risk Framework. Scientific Report DRDC-RDDC-2014-R167

[9] Hellesoy, A. 2014, The world's most mis-understood collaboration toll. [https://cucumber.io/blog/2014/03/03/the-worlds-most-misunderstood-collaboration-tool]. Accessed Jan 25, 2017.

[10] Hansson, A. 2017, Gherkin, [https://github.com/cucumber/cucumber/wiki/Gherkin]. Accessed Jan 25, 2017

[11] Trusted Data Format. XML Data Encoding Specification for Trusted Data Format. https://www.dni.gov/index.php/about/organization/chief-information-officer/trusted-data-format Accessed 27 March 2017.

ANNEX A. CONSULTATION GROUP

Stakeholders		
Name	**Department**	**Stake in Project**
Troy Crosby	ADM(Mat) DGMPD (A&L)	Office of Primary Interest (OPI)
Allan Weldon	ADM(Fin)	Office of Collateral Interest (OCI)
Vrenti Ghergari	ADM(Fin)	Office of Collateral Interest (OCI)
Binyam Solomon	DRDC CORA	Technical Authority
Zakia Bouayed	DRDC CORA	Alternate Technical Authority
Contractors		
Name	**Company**	**Role**
Sandy Lavigne	ISR	Project Manager
Chad Watson	SimFront	Technical Lead
Jason Albert	SimFront	NCR Procurement Specialist
Barry Evans	SimFront	Database Administrator - DBA
Subject Matter Experts		
Name	**Department**	**Subject Matter Expertise**
Kevin Fitzpatrick	ADM(Mat) DGMPD(A&L)	Financial and milestone data
Paula Sams	ADM(Fin)	DRMIS and Costing Data
Chen Hu	ADM(Fin)	Data Requirements for analysis
LCol Kevin Steele	ADM(Fin) CFO	DRMIS and Costing Data
LCol Walton-Simm	ADM Mat	Financial and milestone data
Daniel Hébert	ADM(Mat)	Financial and milestone data
James Sapp	ADM(Mat)	Financial and milestone data
Daniel Perron	ADM(Fin)	DRMIS and Costing Data
Marva George	ADM(Fin)	BI Tool
Trina Paz-Burke	ADM(Fin)	BOBJ Tool
Nancy Desroches	ADM(IM)	Business Warehouse/Business Intelligence
Christian Lachapelle	ADM(Mat)	Senior P & CS Engineer
Paul Desmier	ADM(Mat)	Materiel Group Operational Research (DMGOR)
LCdr Guy Cadrin	ADM(Mat)	Process Control Officer, DGMEPM
Mark Walker	ADM(Mat)	DMPP
Andrew Millson	ADM(Mat)	DMPP
Sue Michaud	ADM(IM)	Solution Architect, DDRMIS

ANNEX B. MILESTONE TIMELINE

Once the scope was set, and the items for analysis were prioritized, a series of tasks and next steps naturally emerged. The timeline and results are noted below:

Jan 25th – Meeting (Stakeholder Session 1) - Scoping

Jan 25th Session 1 was characterized by scoping and resulted in:

- Scope 2d-2j-2a established
- Data Sources revealed:
  - CID
  - DRMIS
  - SME Opinion (MEOSAR)
- new SMEs in Working Group Paper
  - Paul Desmier and
  - LCdr Cadrin added.
- Initial Glossary
- Follow on analysis tasks

These results led to a series of analysis tasks leading into the next stakeholder session.

Feb 1st – Predicting Project Milestones:

- Initial Executable Specs
- First exposure to the “Business Flow”
- Logical Data Concepts

Feb 3rd – DRMIS and CID demos:

- SWOT analysis
- Data Dictionary updates

Feb 9th – Business Flow

- Initial concept of environment

Feb 13th – Meeting (Stakeholder Session 2) – Ambiguity

A concept for a historical costing database was prepared for session number two. Consensus was expected, but not achieved. A significant amount of discussion emerged in relation to the definition of cost, particularly in regards to definitions that are found in other documents (previously unknown) such as a cost breakdown structure (CBS).

Feb 13th Session 2 was characterized by ambiguity and resulted in:

- Logical Data Concepts – Un-expected ambiguity in relation to the definition of a “cost”.

A new concept was required. Another stakeholder session was required. The topic of CBS was raised and triggered significant discussion about costs. For example, can a milestone have a cost that is estimatable?

- Business flow – consensus and a more refined picture of the environment – several updates.

These results led to series of analysis tasks leading into the next stakeholder session.

Feb 22 – HEAP – Historical Estimates and Actuals Program TA Review

- Design for Remaining Ambiguity
- Concept of a Risk Information Exchange Data Model

Feb 23 – "2j"

- Ability to capture estimates through "Plan Versions" in "Project Systems"
- DRMIS has a lot of capability in it, but you need to know what to ask for. Now we know to ask for "Plan Versions"

Mar 2 – HEAP – Sample SME Review

- Confirmation that the model could deal with the remaining ambiguity

Mar 8th – Meeting (Stakeholder Session 3) – HEAP

The concept for HEAP was briefed to members of the consultation group.