Version <>

[Note: The following template is provided for use with the Rational Unified Process. Text enclosed in square brackets and displayed in blue italics (style=InfoBlue) is included to provide guidance to the author and should be deleted before publishing the document. A paragraph entered following this style will automatically be set to normal (style=Body Text).]

**[Note: This version of the template has updated the version number to read from metadata (Properties->Custom->Version) and the header date to read from metadata (Properties->Custom->Version Date).**]

[To customize automatic fields in Microsoft Word (which display a gray background when selected), select File>Properties and replace the Title, Subject and Company fields with the appropriate information for this document. After closing the dialog, automatic fields may be updated throughout the document by selecting Edit>Select All (or Ctrl-A) and pressing F9, or simply click on the field and press F9. This must be done separately for Headers and Footers. Alt-F9 will toggle between displaying the field names and the field contents. See Word help for more information on working with fields.]

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Version** | **Description** | **Author** |
| 3/26/2012 | 0.1 | Initial version | Simon Chatwin |
| 4/3/2012 | 0.2 | Feedback from Ying | Ying/Simon |
|  |  |  |  |
|  |  |  |  |

Table of Contents

[1. Introduction 3](#_Toc321329563)

[1.1 Scope 3](#_Toc321329564)

[1.2 Definitions, Acronyms, and Abbreviations 3](#_Toc321329565)

[1.3 References 3](#_Toc321329566)

[2. Stakeholder and User Descriptions 3](#_Toc321329567)

[2.1 User Summary 3](#_Toc321329568)

[2.2 User Environment 4](#_Toc321329569)

[2.3 User Profiles 5](#_Toc321329570)

[2.3.1 Assay Registrar 5](#_Toc321329571)

[2.3.2 Dictionary Curator 5](#_Toc321329572)

[2.3.3 Data Depositor 6](#_Toc321329573)

[2.3.4 Data Analyzer 6](#_Toc321329574)

[2.4 Persona Stories 7](#_Toc321329575)

[2.4.1 Curator 7](#_Toc321329576)

[2.4.2 Registrar 7](#_Toc321329577)

[2.4.3 Depositor 7](#_Toc321329578)

[2.4.4 Analyst 7](#_Toc321329579)

[2.5 Key User Needs 7](#_Toc321329580)

[3. User Workflows 8](#_Toc321329581)

[3.1 Assay definition and approval 8](#_Toc321329582)

[3.2 Data Upload 8](#_Toc321329583)

[3.3 Dictionary curation 8](#_Toc321329584)

[4. Features 8](#_Toc321329585)

[4.1 <aFeature> 8](#_Toc321329586)

[4.2 <anotherFeature> 8](#_Toc321329587)

# Introduction

The purpose of this document is to collect, analyze, and define high-level needs and features of the BARD CAP and data entry system. It focuses on the users, their needs and workflow processes, and **why** these needs exist.

## Scope

This document is limited to the collection of data for the BARD system. Other docuemnts describe ghe uses of the data and the following analysis that is performed.

## Definitions, Acronyms, and Abbreviations

|  |  |
| --- | --- |
| MLP | Molecular Library Program – an NIH financed project to create and characterize libraries of small molecules |
| Dictionary | Data Dictionary or ontology. In this document these are taken to be the same thing, though their implementation and capabilities may be very different. It is a store for terms that can be used in an assay definition |
| Assay | The set of result types, conditions, and design parameters that define the results in their scientific context. An assay has no substances attached to it (unlike PubChem – see “Experiment” below). |
| Experiment | This is an assay run or a collection of results from several assay runs gathered together for data entry in one session. The Experiment more closely mirrors the PubChem definition of an “Assay”. Also known as an ‘Assay Instance’ |

## References

<TBD>

# Stakeholder and User Descriptions

## User Summary

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Description** | **Knowledge requirements** | **Stakeholder** |
| Current Roles at work | | | |
| Biologists | Scientist who develops the assay, runs it and calculates results. Interested in doing the minimum necessary to ensure grant money keeps flowing | Deep understanding of the biological parameters and conditions for an assay |  |
| Informatics Specialists | Experts in IT systems who support the biologists and take the task of converting data into suitable format for upload. Has a broader understanding of data needs. | Knowledge of the needs and consequences of using a data system. Enough understanding of the biology to enhance communication with Biologists |  |
| Curators | Senior biologist in PubChem who approves assay definitions and results | Knowledge of PubChem and its contents together with the restrictions and consequences of entering data in particular ways |  |
| Chemist | Chemists are generally data consumers. | Knowledge of chemical structures and their diagrammatic representation. |  |
| Chemin-formatics Specialist | Compound deposition is done by Cheminformatics specialist. |  |  |
| New Roles in BARD – these not necessarily map to the roles above and one person may have more than one role | | | |
| Curator | Manage and approve the changes and updates to the data dictionary/ ontology | Requires comprehensive knowledge of biology and biological terms.  Requires knowledge of how BARD operates and the consequences of setting up the dictionary in a particular way | Not identified |
| Assay Registrar | Registers assay designs in BARD | Requires knowledge of the biology used in the assay and what might make one assay distinct from another | Not identified |
| Data Depositor | Uploads data from experiments into BARD using a pre-defined assay definition | Requires knowledge of how BARD stores data and how results and their contexts depend on the assay definition | Not identified |
| Data Consumers | Queries data from the system | Knowledge of biological assays | Not identified |

## User Environment

There are currently about 15 MLP centers contributing biochemical data to PubChem. Each of these centers uses different techniques, processes and tools to define assays and load them into PubChem. The people involved usually hold several roles (e,g, curator and assay registrar) and currently most come more from an informatics background than a biology background. The current workflows for loading data into PubChem are considered intrusive and not suitable for use by biologists (the complexity takes too long to understand and use with little benefit to the user).

In general the benefit of good data entry (high quality data) is not received by the entry personnel but by the data query experts. This separation of workload/effort and benefit makes persuasion to encourage data entry difficult.

Users are not generally IT sophisticates.

## 

## User Profiles

| Number and name | Representative | Type | Responsibilities | Success Criteria | Involvement | Deliverables | Comment /Issues |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Assay Registrar |  | Biology expert with little interest in IT.  Several of these may exist at each MLP center. | Create and approves assay definitions, proposes new dictionary elements for the dictionary. Will educate other biologists on assay registration and ask biologists to provide the necessary information to register the assay. | Registering an assay takes a sufficiently small time that it is not worth delegating to an informatics professional.  Generally data entry is rewarded by continuation of grant funding. Data quality does not matter. | Requirements provider and UAT tester. | Requires rules for what is the minimum data required for an assay definition | This user very rarely makes queries of the system and does not feel the need f or quality data. He bears the brunt of the data entry workload |
| Dictionary Curator |  | Biology expert with knowledge of hierarchies and ontology. Reasonable familiarity with contents and usage of the dictionary | Makes entries in and cleans up the dictionary. Reports to the RDM team about the state, completeness and accuracy of the dictionary | General management of new terms takes less than 10 mins per term.  Not involved in assay registration and approvals | Requirements provider and UAT tester.  Determines the contents of the dictionary during data migration.  Assists in work of annotating assays for data migration | Contents of the data dictionary | This person is primarily responsible for the quality of the data in BARD. While this role may overlap with the Analyzer, the curator rarely makes queries of the system |
| Data Depositor |  | Some biology knowledge and good understanding of the formats, limits and consequences of BARD data entry.  Expert Excel user (for massaging data before upload) | Retrieves the data format for the assay definition, prepares the data for upload (converting formats etc.).  Will educate biologists and others about the data needs of BARD to ensure good quality data is received for upload | Minimal interaction with the curator to get data loaded.  Time to load data for an experiment | Requirements provider and UAT tester.  Heavily involved in the Data Migration effort | High quality data extracted from PubChem and loaded into BARD | Often this person comes from an informatics skill set so that handling of the large amounts (up to 1M rows) of data can be done with appropriate tools |
| Data Analyzer |  | TBD | TBD | TBD | TBD |  |  |

## 

## Persona Stories

### Curator

The curator may be a senior researcher with an interest in ontologies and a broad scope of biological knowledge based on terms in the ontologies they manage.

The curator carries the highest responsibility for good quality data in the system, though he may not often be a consumer of the data. Demands on the curators time are extreme and so curation must be a quick and seldom used task. Curators are good at using automation to achieve a lot of the integrity checking when adding terms to an ontology

### Registrar

The assay registrar is an experienced biologist with a PhD and several years of leading research teams. His work is defined by the grants s/he has won and often his/her career depends on the continued satisfactory performance to the terms of the grant. Thus uploading of data into a public resource is regarded as a necessary nuisance that gets in the way of day-to-day goals. The registrar is rarely a consumer of data and so has little interest in the data quality and how it is received by the Analysts; his/her primary measure of quality is in the peer-reviewed publications and the associated citations.

His favorite tools are usually Excel and some Powerpoint and Word. Many of the more specialist tools that data comes from are actually operated by junior personnel who report up to him.

Each data upload center (laboratory) will have several registrars and they may delegate many tasks perceived as ‘IT’ tasks to others.

### Depositor

The depositor is usually a more junior biologist or may be an Informatics person. This person spends much of their time at a computer finding data, combining and evaluating it and presenting it for use by senior biologists (see Registrar). They are very familiar with Excel and use many other specialized IT tools in their daily tasks.

A significant portion of their time is spent in the laboratory, but even this involves much work setting up experiments to be performed robotically.

### Analyst

## Key User Needs

|  |  |  |  |
| --- | --- | --- | --- |
| **Need** | **Priority** | **Concerns** | **Proposed Solutions** |
| Simple operation with no need to relearn the system |  | Often uploads are done several months apart and users forget the business rules | System must provide guidance to users.  System must implement business rules in as automatic a fashion as possible |
| Assay definitions are not entered until results are ready to upload |  | Users don’t want to wait for dictionary elements to be approved before uploading experiment results | Allow ‘pending’ entries to the dictionary to be used for result upload.  Data can be adjusted after the elements are approved. |
| Efficient processes for approving results |  | Users want to be able to get this chore [of loading data] over with | Automatically approve results if the assay is approved. If the assay is approved later due to dictionary issues, the results are automatically approved as well. |
| Encourage communication between approvers and enterers of data |  | Since a user may have infrequent access, learning opportunities must be maximized by ensuring that errors in data entry are corrected by the data enterer, not the curator/approver | On-line Help  Simple ways to contact the experts (chat?) |

# User Workflows

See also the [UML Model](http://broadinstitute.github.com/BARD/CAP_and_Data_Entry/) in the GitHub Wiki pages

## Assay definition and approval

## Data Upload

## Dictionary curation

<TBD>

# Features

Features are listed in the Pivotal Tracker tool where the SCRUM backlog is maintained. That list is dynamic and well maintained. This list is the higher level features that are expected to be stable during the initial implementation of the system

## Dictionary and terminology Source

The data dictionary driving thr contents of the assay definitions is maintained outside the suite of tools and will be imported by an .obo file from an OWL ontology.

The dictionary divides into several areas (major categories) that are used in different areas in the CAP, data entry and Assay management tools.

## Data entry is seamless

Most centers combine the roles of Registrar and Depositor so they tend to combine the two functions.