Initial Thoughts on the Prices Paid Project

These are our initial and evolving thoughts on the Prices Paid project.

In addition to the documentation which Laura Eyester has already provided, we make the following preliminary assumptions:

1. We want to make it easy for agencies to report data to us, even though they may be required to do so by law. Since every agency may have different formats or none at all, we want to facillitate the harmonizing of this data.
2. Price data may be valuable to a use even if it is not yet harmonized, if they are able to search it and understand the limitations of the non-standardized data.
3. Some agencies and some kinds or purchased services and products will be easier to harmonize than others.
4. The harmonization of pricing data will necessarily evolve over a long time.
5. We want to present an read-only Reporting API to the data that can be used by anyone.
6. We want to make it easy for the RFP-EZ users to get access to the prices paid information via a GUI of some kind.
7. Harmonized price data can be presented in a graphical form. Free-form data can be viewed in a google-like search form.
8. We will want to be able to upload data sets and retain metadata control over those data sets. For example, if a data set is discovered to be erroneous, we want to be able to mark it as such or to remove it. Similarly, we want users to be able to test uploads by marking a dataset as “test” and limiting who can see it until they remove the “test” setting.

Based on these unvetted assumptions, we have designed a preliminary architecture, show below.

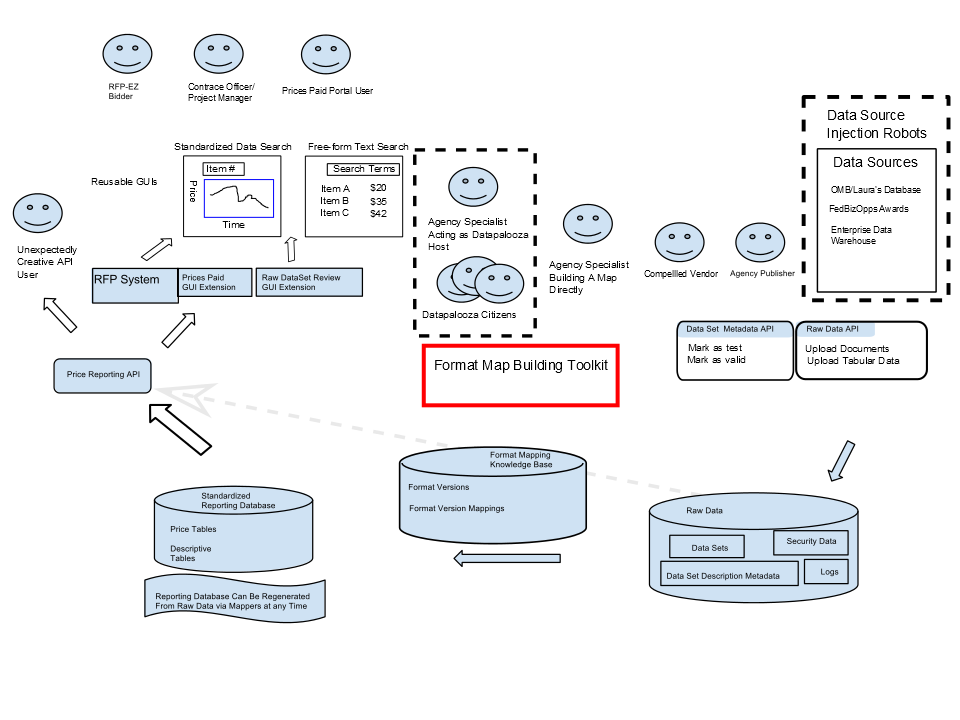


Figure : Data/Work Flow of Project

The basic approach the diagram is trying to show is to separate the “Raw Data” database from the “Standardized Reporting Database (SRD).” The Standardized Reporting Database can be created regenerated from the Raw Data at any time by using the third database, which is a set of mappings from Raw Data formats to standard data formats which we have yet to create. However, we can provide APIs and GUIs to search both the Raw Data and the SRD.

We imagine that RFP-EZ system will be extended to use the search APIs, so that an RFP-EZ user can have access to the prices paid data.

# On the Format Mapping Knowledge Base

The Format Mapping Knowledge Base serves three purposes:

1. It catalogs, tracks, and describes various data formats that contain pricing information that may come from diverse agencies.
2. Its mappings can be “executed” to convert the raw data into the Standardized Reporting Database.
3. It serves as a model and knowledge base for how to construct mappings between the various formats.

We believe that the creation of the mappings between various formats is a tedious, time-consuming, and ever-evolving task that often will not have a clear or perfect solution. The total labor to create all of the mappings between raw data formats and the Standardized Reporting Database might be One MILLION person-hours. The only way to attack a problem of that size is to organized disciplined “data jams” that allow a form a crowd-sourcing of the task of the harmonizing data formats. The “crowd” could be government employees who are specialists in the data formats, government employees more generally, or even the public at large under some conditions.

Before such a “Harmonizing Data Jam” can be created, we will have to have:

1. A mechanism for defining the data mapping,
2. A mechanism for uploading the data mapping into the database,
3. “Data Marshalls” to review the mappings and test the results,
4. An ability to “execute” the data mapping to usefully improve the data that can be viewed via a GUI hooked to the Price Reporting API.

# Sources of Data

We are attempting to understand various sources of data. These include:

* The FedBid data
* The Enterprise Data Warehouse, which compiles 4 systems:
  + Advantage [EBuy]
  + OS2
  + ReverseAuction
  + MITCB(sp?)
  + SSSIT (SSS19)?

## A GUI Design

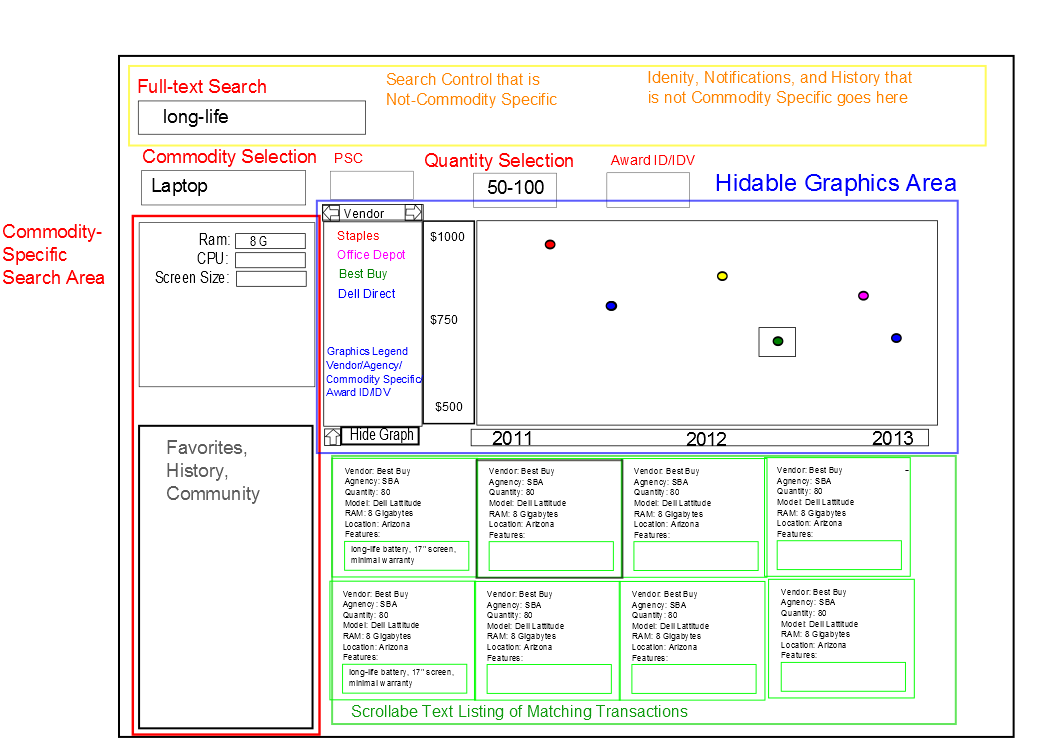


Figure : Prices Paid GUI Design

Elements of this GUI Design was suggested by Laura Eyester based on the Amazon.com user experience. The most important points are:

* The user sees and interacts first with google-like Full-Text Search. This is used to find and limit transactions which are displayed in the Blue and Green areas.
* The full-text search may suggest Commodities. Once a Commodity is chosen, the Red area is populated with Commodity specific data fields. These can be used to both understand the data and to further limit the search based on commodity-specific choices.
* Also in the yellow area are selections for the PSC, Quantity range to be studied, and perhaps other fields that limit the search regardless of commodity type.
* The blue area is a hidable graphics area. This supports 508 compliance and those users who prefer to see the data textually. When the graphics are hidden, more screen space is used for the Green area, allowing more postings to be shown.
* The Green area shows individual transactions in a textual way. The will be ordered by a criteria not shown in the sketch. A selected transaction can be highlighted in some way, shows as a box in the figure.
* The Blue area is a graphic display that shows a dot for each transaction matching the current search criteria. The graph gives a visual indication of unit-price over time. The color is used to map to an additional piece of data, such as Vendor Name, chosen from the set of all data fields.