Phase One of the Comprehensive Extensible Data Documentation and Access Repository

Jeremy Williams, William Block and Warren Brown, Cornell University

Project Background

The United States federal statistical system produces prodigious amounts of both public and confidential data valuable to social science and economic research. These data can be difficult to interact with due to poor documentation. In the scientific community, results must be reproducible to be proven accurate. This condition is hindered without proper metadata to describe the original context and intent of a given study. In recognition of this, the National Science Foundation has recently instituted a requirement for all grant proposals to include a plan for data generated during research to be persisted and made available for future research.

Many organizations have also set out to establish metadata to describe the rich resources of the federal statistical system. This introduces another layer of complexity as these organizations do not all adhere to a single standard, making the metadata useful, but insular. The Data Documentation Initiative (DDI) is an emerging metadata standard that is used internationally to describe data in the social sciences. It has the potential to unify the metadata managed by separate organizations into a comprehensive searchable set.

Researchers from the Labor Dynamics Institute, in collaboration with the Cornell Institute for Social and Economic Research (CISER) received funding from the National Science Foundation to improve the documentation of US federal statistical system data with the goal of making it more discoverable, accessible and understandable for scientific research. The project has been named CED^{2}AR (Comprehensive Extensible Data Documentation and Access Repository). The objective for CED^{2}AR is to provide a facility to make standardized metadata from heterogeneous sources searchable through an online interface designed to be intuitive to researchers.

Phase One of the Project

Phase One, a subset of the overall CED^{2}AR project, is to develop the search API and the web interface for user searches. This presentation of Phase One covers the following components: project background, deliverables, technical feasibility, system requirements, system and program design, user interface design, and user testing. Some these components are illustrated in this proposal.

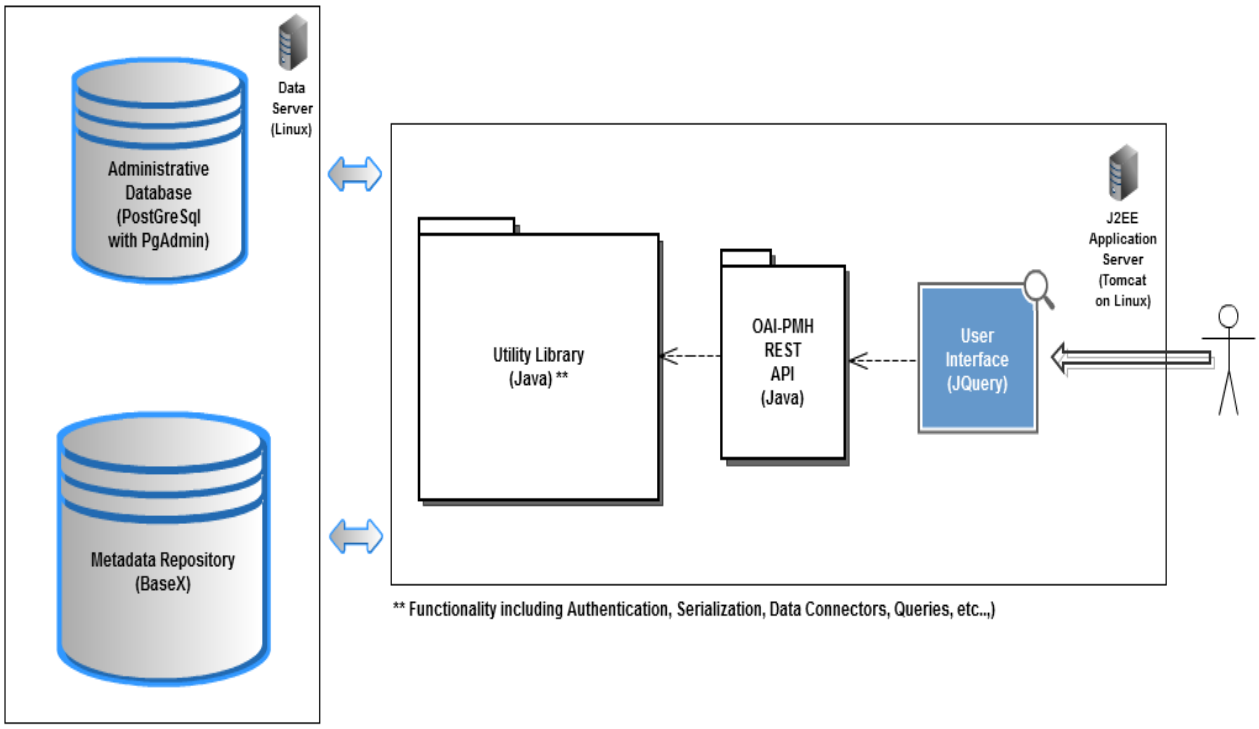
Deliverables

- *A fully functioning web interface for users.*
  - The users can search the metadata database
  - Data search filter should be available before and after the results being displayed.
  - The users can sign up for an account.
  - Users should be able to have access to their account using security questions and they should have the ability to change their password.

- *A well documented API that returns data in XML, JSON and CSV.* This API is another core work-product of this project. This API should be well documented so other programmers can use it without much training. It must be built upon standards (such as HTTP, REST, OAI-PMH) and support multiple representations of the resources of the repository (JSON, XML, CSV, etc..,). It must also implement the idea of connectedness, allowing resources to be linked to related resources for easier navigation and resolution.

Technical Feasibility

Shown below is a diagram of the system design for Phase One of the project.



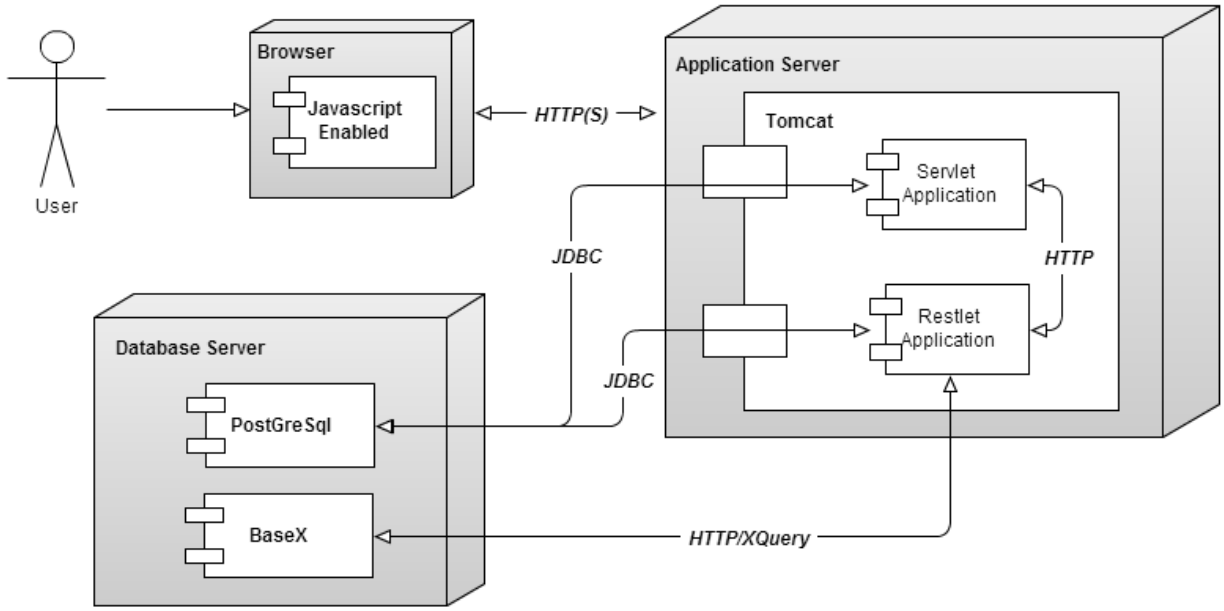
The metadata repository and the administrative database are on the left side of the diagram. The project encompasses the user interface and API seen on the right side of the diagram. The process of ingesting metadata from various sources into the repository is (and has been) underway, but is outside the scope of this project.

System Requirements

There are three main components related to our project: a metadata repository, an application programming interface (API), and a web interface. Phase One of our project implemented the API and the web interface. At all levels of the project, complying with standards and providing sufficient documentation were of the utmost importance in order for the end users to be able to easily operate the final deliverable. Since the interface is a web-based application, accessibility and security were also considered throughout the development process.

System Design

Multi-Tier Architecture



The diagram above shows a modular representation of the system design. Each box represents a logical tier (including client, application, and database). Within each box resides a component which is connected to one or more dependencies by the protocol indicated.

Data Model

The primary data model being utilized in this application is DDI 2.5. The system will also use a simple relational database to handle user interactions with the site, such as logging in, and saving queries.

Application Programming Interface

The CED2AR API serves as the intermediary between the user interface and the repository's XQuery Processor. It takes a specified set of user generated inputs over HTTP and interprets them into the corresponding XQuery text. This text is passed to the CED2AR repository where it is processed. The results of the query are then passed back to the user interface over HTTP in the representation format requested (e.g. XML, JSON, CSV, etc..,). This includes both searching and exposure of all data in the repository.