Dataset Support in SobekCM

# Overview

Software Developers for the SobekCM Open Source Software[[1]](#footnote-1) have begun adding data support for specific data archiving (with versioning) needs, and with support for queries, searches, sorts, report generation, and other actions with archived data.

# About SobekCM

The SobekCM software is a full suite of applications that power digital libraries, digital content/asset management, digital preservation, discoverability, online patron user tools, and workflow tools for integration with library and other web-scale systems, digital production, and digital curation. SobekCM is the software engine which powers many digital libraries, exhibits, digital production workflows, and more at institutions around the world including the Digital Library of the Caribbean (dLOC), Florida Digital Newspaper Library, the University of Florida Digital Collections (UFDC), and many others.

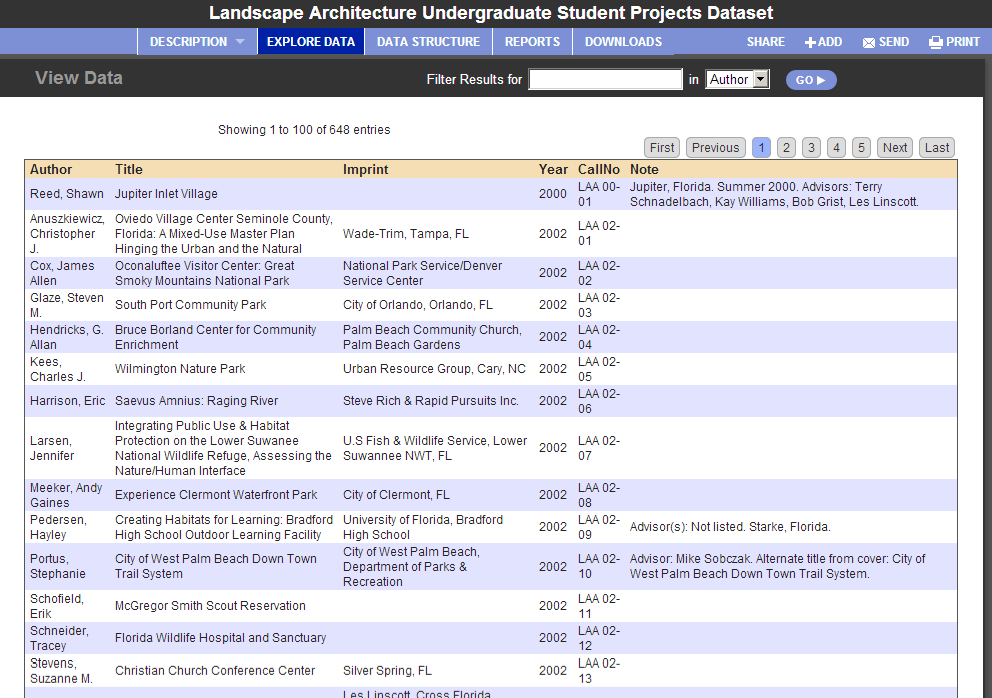
SobekCM allows users to discover online resources via semantic and full-text searches, as well as a variety of different browse mechanisms. For each digital resource in the repository there are a plethora of display options, which may be selected by an appropriately authenticated use. This repository includes online metadata editing and online submissions in support of institutional repositories.

# Dataset Support in SobekCM: Prototype Development & Work to Date

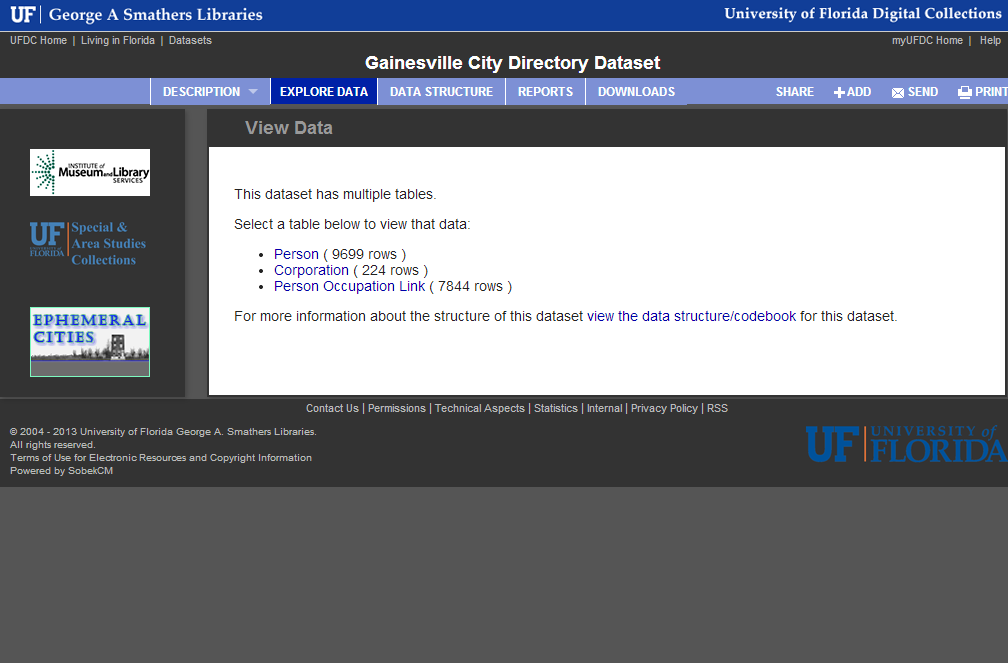
In October 2013, the Development Team for SobekCM at UF, led by Mark Sullivan, began adding prototype dataset support to the *Institutional Repository @ UF* (*IR@UF*).

* Prototype for final display of datasets (dataset with a single datatable): <http://ufdc.ufl.edu/IR00003504/00001>
* Prototype for dataset with multiple tables:   <http://ufdc.ufl.edu/AA00017819/00001>
* Prototype for dataset codebook: <http://ufdc.ufl.edu/AA00017819/00001/dscodebook>

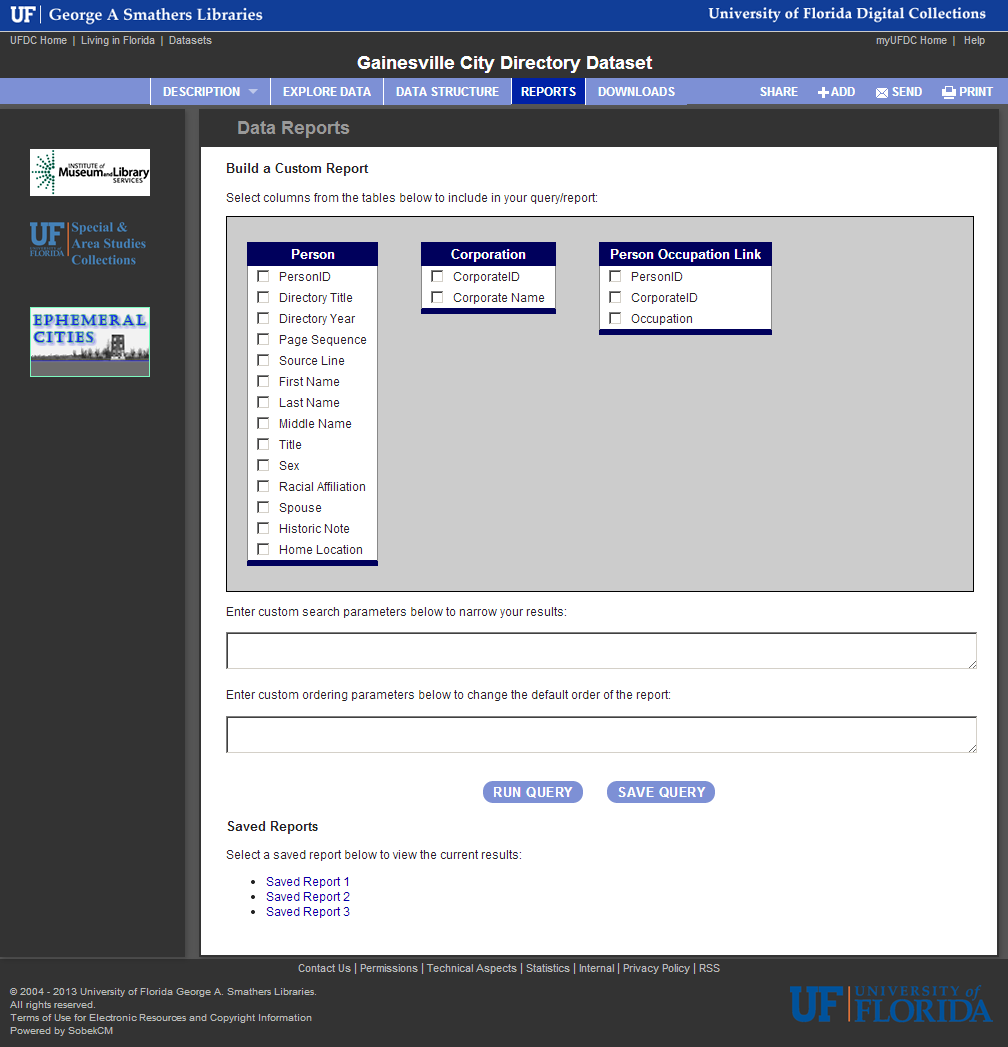
## Prototype for final display of datasets (dataset with a single datatable)

****

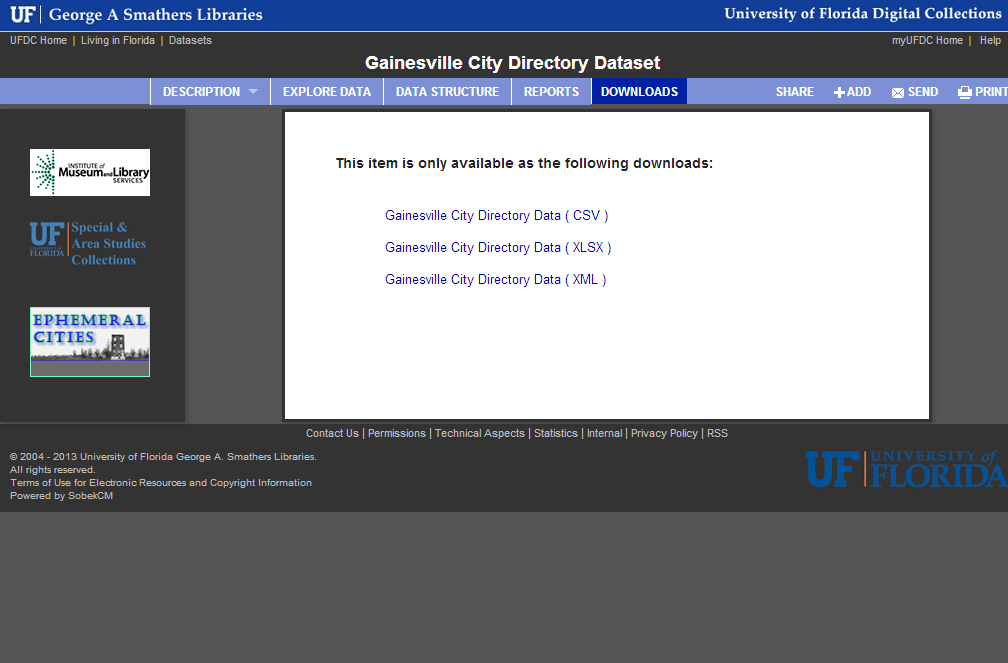
## Prototype for dataset with multiple tables

****

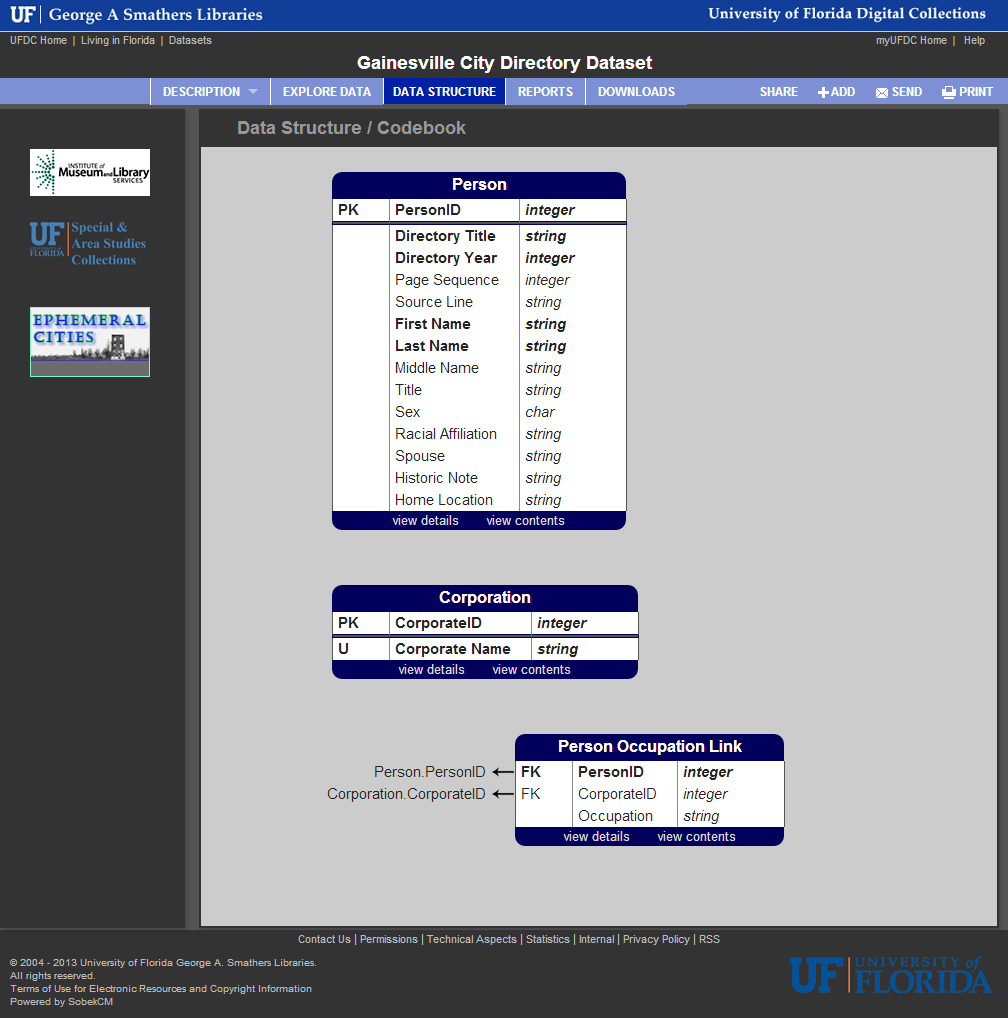
## Prototype for dataset with multiple tables: reports

****

## Prototype for dataset with multiple tables: downloads

****

## Prototype for dataset with multiple tables; dataset codebook

****

In these examples, everything (e.g., the code book, uniqueness and foreign key constraints, required fields, etc.) are derived from the XML schema included at the top of the XML.  The XML schema is viewable under the “downloads” link.  The schema currently uses Microsoft as the extension schema, which is the first support with more to be added.   
  
Paging through the data in particular is powered by a back-end data provider which serves JSON to the jQuery datatable plug-in.  This is becoming more of an interface norm for data services on user-focused enterprise service sites, and provides a familiar framework for users here for the data services. This interface will likely continue to be written in JavaScript/jQuery that reads JSON to draw the tables for example.  Currently, the prototype HTML is written directly in C# code.

Considerations for the IR@UF Presentation

Clearly, a major part of the problem is normalizing Excel and CSV files into XML and retrieving information from the user about each row, how it should be searched, etc.  
  
Clicking on a single row doesn't retrieve the correct row, nor does it yet travel through the table relations to show information in related tables.  For the interface as currently envisioned, this screen will include a button for “edit this row” and a button on the main view data screens to “add a row.”  The input/edit forms are expected to be created directly from the XSD's information.  
  
The prototype is currently working with a XML NoSQL solution, but using the XSD solution with a SQL back-end should work well.  The system could parse the XSD to discover the structure/codebook and everything else would be relatively similar.  Instead of retrieving the data from the dataset derived from XML, it would be read into a dataset from SQL.  The one difference is that only the data needed immediately for display would be retrieved from SQL (probably with paging through the data for handling big data).

## Considerations for System Integration with the Libraries, Research Computing, and/or Others

In addition to considerations for the IR@UF presentation and functionality as powered by SobekCM, this SQL could reside on servers supported by the UF Libraries, Research Computing, and/or others.  With additional collaborative development, the back-end would be able to use Hadoop or iRODS and the additional development would enable it for “Big Data” presentation.

# Dataset Support in SobekCM: Next Phase

At this time, Mark Sullivan (Application Engineer for SobekCM; Head of the UF Libraries’ Digital Development & Web Services Team with many SobekCM Developers) is planning to pursue an Emerging Technology Grant from within the Libraries, seeking $10,000 for developer salary. At this time, one of the developers on the Digital Development & Web Services Team is on grant funding, with a time gap between when the current grant project funding ends and the next begins. This presents the opportunity to immediately hire a skilled full-time developer for a specific project and defined timeframe, and Dataset support in SobekCM has been selected as the appropriate project for this period.

The proposed project for the Emerging Technology Grant will focus on adding support for simple visualizations including simple graphs and mapping, possibly similar to CKAN.[[2]](#footnote-2)

Dataset Support in SobekCM: Immediate Next Steps

For the immediate future, Mark Sullivan is working towards the grant proposal and continuing to refine the dataset support in SobekCM as time allows. He will share updates on progress and any considerations for discussion with UF’s Campus-wide Data Management/Curation Task Force[[3]](#footnote-3) and the other SobekCM Developers.[[4]](#footnote-4)

# Document Information

*Dataset Support in SobekCM* (Oct. 2013). Technical information written by Mark V. Sullivan, initial text revised and expanded for use as documentation and news by Laurie N. Taylor.

1. SobekCM is released as open source software under the GNU GPL license, and is downloadable from the SobekCM Software Download Site, which also has complete information and technical and code documentation: <http://sobek.ufl.edu/> [↑](#footnote-ref-1)
2. <http://demo.ckan.org/dataset/gold-prices/resource/b9aae52b-b082-4159-b46f-7bb9c158d013> [↑](#footnote-ref-2)
3. <http://library.ufl.edu/datamgmt/> [↑](#footnote-ref-3)
4. <https://groups.google.com/forum/#!categories/sobekcm-discuss/general> [↑](#footnote-ref-4)