IMLS National Leadership Grant:
LG-05-10-0104-10
Built Works Registry (BWR)

Interim Narrative Report
September 30, 2011 – October 1, 2012

Date: October 31, 2012
Project Director: Carole Ann Fabian
Director, Avery Architectural & Fine Arts Library
212-854-3068
caf2140@columbia.edu

Prepared by: Avery: Carole Ann Fabian
ARTstor: James Shulman, Gretchen Wagner, Bill Ying

Interim Report Summary

In this second year of IMLS grant funding, the collaborative efforts of the project's teams at the Columbia University Avery Architectural & Fine Arts Library, ARTstor, and the Getty Research Institute (GRI) have focused on implementation of the technological infrastructure, data analysis of seed content, and development of a geo-coding strategy for the Built Works Registry (BWR). Institutional data contribution policies drafted during the first year have been finalized and the BWR administrative team is working with counsel at data-providing institutions to secure an open, data-sharing agreement. The following progress report summarizes in greater detail progress in each of the areas outlined in the project timeline (Addendum A).

Project Administration

Staff:

Staffing established during the first year continued through the second with few changes. Project staff replacements and/or additions occurred in one partner organization, but this staff change has not affected budget allocations:

- Avery Library – no changes in staff
- ARTstor: Christine Kuan left ARTstor and is replaced by Giovanni Garcia-Fenech (Communications Manager). See Addenda A: Staff CV.
- Getty Research Institute – no changes in staff

Team meetings:

Staff from the three partner organizations participated in BWR all-team meetings, which alternated between ARTstor and Avery Library in New York, with Getty members taking part via conference call or webinar. Additionally, each BWR project team (administration, metadata, technology) convened regular meetings and engaged in ongoing email/phone communications to ensure progress of work. The Built Works Registry metadata team also met with the ARTstor

Shared Shelf and Getty Vocabularies teams to ensure that data development would interoperate across all systems.

BWR Advisory Council meeting:

An onsite BWR Advisory Council *http://builtworksregistry.wordpress.com/advisory-board-members/* meeting was held at ARTstor on February 3, 2012; members attending included:

Bilson, Tom (Courtauld Institute of Art, London)
Blair, Sheila (Boston College, Boston)
Bloom, Jonathan M. (Boston College, Boston)
Chiku, Kakugyo S. (Kanazawa Institute of Technology, Kanazawa)
Cope, Aaron Straup (Cooper-Hewitt Design Museum, New York)
Farneth, David (Getty Research Institute, Los Angeles)
Hagedorn-Saupe, Monika (Institute for Museum Research, Berlin)
Murray, Irena (Royal Institute of British Architects, London)
Quigley, Sam (Art Institute of Chicago)
Saliga, Pauline (Society of Architectural Historians, Chicago)
Simane, Jan (Kunsthistorisches Institute, Florence)
Sinha, Vandana (American Institute of Indian Studies, Gurgaon, Haryana, India)
Stubbs, John (World Monuments Fund & Tulane University, New Orleans)
Waibel, Günter (Smithsonian Institution)
Whiteside, Ann (Harvard University Graduate School of Design & Frances Loeb Library)

Together with thirteen BWR staff participants, a full-day program covered the following topics:

- BWR continuity & sustainability planning:
  Proposed terms of the BWR Founders Agreement**,** BWR Institutional Contribution Agreement**,** and BWR Terms of Use.

- Shared Shelf platform & BWR technical infrastructure:
  Overview and system demonstration

- BWR metadata schema:
  Overview and discussion of brief and extended metadata schema; correlation with Getty Cultural Objects Name Authority schema.

- BWR Tools:
  Discussion of potential use case scenarios to assist in formulating functional requirements for data contribution environments. Discussion of challenges in aggregating large data sets from institutional records.

- BWR data strategies:
  Overview of proposed approach for working with the five seed data sets including: disambiguation, prioritization, and normalization.

- BWR policies for collaboration:
  Presentation and discussion of other large-scale data aggregation projects including VIAF & the International Terminology Working Group and the Society of Architectural Historians' SAHARA editorial model.

- Looking ahead:
  Discussion of future collaborations with institutions and contributions by individuals; current and future role of the BWR Advisory Council .

Dissemination:

Information about the Built Works Registry project is regularly communicated to relevant constituencies in a variety of presentations, print and electronic media. The Built Works Registry blog *http://builtworksregistry.wordpress.com/* continues to serve as the main dissemination venue for the project; postings throughout the year provide updates on project activities.

Avery Library and ARTstor staff presented updates on the BWR project at the following professional conferences:

- Association of Architecture School Librarians (Boston, March 2, 2012)
- WebWise 2012 (Baltimore, March 2, 2012)
- ARLIS/NA Conference (Toronto, March 30, 2012)
- Visual Resources Association Annual Meeting (Albuquerque, April 19-20, 2012)

Avery and ARTstor staff are developing a range of marketing materials to demonstrate the value of BWR to institutions, including a number of case studies based on the experiences of existing contributors. We are anticipating a new prospective contributor webinar, which will focus on different use cases, targeting catalogers, librarians, and institutional contributors.

Policies

We are currently in the process of working through the agreements for the seed collections (with institutional counsel) and anticipate that this process shall be completed shortly. Cornell and Harvard have indicated in principle that they are fine with proceeding, and their counsel have copies of the proposed content contribution licenses. Both institutions are trying to determine whether they will move forward using the license we proposed or a CC Zero license.

Content

BWR will identify, extract, & verify data from five sources: Harvard University's OLIVIA database, Cornell University Libraries' PICTOR database, the *Avery Index to Architectural Periodicals*, Avery Library's AVIADOR database, and selected architecturally-focused collections from the ARTstor Digital Library database. In each case the seed data sets have been evaluated for relevant content and core records are being extracted as follows:

- Harvard's Olivia 2.5M > 392,917 work records > **55,621** unique sites

- Cornell's Pictor 24K > 13,847 records > **1,611** unique sites
- Avery's AVIADOR 40,000 records > **4,878** unique sites
- Avery Index 650,000 records containing **~433,000** subject term strings
- ARTstor 1.4M records containing **~70,000** built work image records

As expected, while each data set contains records for many of the same built works, the records are non-standard and diverge from each other significantly in most data values. The following chart demonstrates a typical array of data for Frank Lloyd Wright's masterwork, Fallingwater:

Field Name	AVIADOR	Harvard	Cornell	Avery	ARTstor
Title/Name of Building	Fallingwater, Kaufmann House	Fallingwater, Edgar J. Kaufmann House, Kaufmann House	Falling Water, Edgar J. Kaufmann Residence	Fallingwater (Kaufmann house).	Fallingwater, Edgar J. Kaufmann House, Kaufmann House
Location	Bear Run, (Pa.)	Bear Run, Pennsylvania, United States	Bear Run, PA, United States, North and Central America	Bear Run (Pennsylvania), United States	Mill Run, Pennsylvania, United States
Creator	Wright, Frank Lloyd One record with Ferriss, Hugh	Wright, Frank Lloyd	Wright, Frank Lloyd	Wright, Frank Lloyd	Wright, Frank Lloyd
Date		1935-1939	1934-1938		1935-1936
Work Type	Vacation houses, Houses, Country Houses	houses	House		House
Style/Period			Modern		Modernist
Subject		Summer houses; vacation houses; terrace houses; hillside architecture	Houses		Dwellings, Houses, Cantilevers, Waterfalls, Architecture—United States, National Historic Landmarks Program (U.S.), Architecture, Modern—20th century, Kaufmann, Edgar J., 1885-1955
Culture		American	American		
Description			Modern Architecture		
Other names		Kaufmann, Edgar J. (patron)			

Technology & Metadata

Since the last report, progress has been made in preparing the technological infrastructure to support the Built Works Registry in a number of inter-related areas: the server/network infrastructure; the data model and data warehouse; the data merging tools; the record building and record retrieving interface; and the geo-coding strategy.

Server/network infrastructure

The infrastructure that will support the data warehouse and the record building/record retrieving interfaces are hosted in our co-location server farm in downtown Manhattan. Co-hosted with the Shared Shelf service, the BWR infrastructure and data warehouse have dependable disc storage, processing power, network, and authorization and access capacities for the foreseeable future.

Data model and data warehouse

In year 1 we established the core and extended data schema and data dictionary. This schema forms the common framework for all institutional and individual data contributions. Data constructed using this schema will be ingested into a repository called the Data Warehouse. The Data Warehouse will contain all the contributed registry work records as well as ARTstor's Name Authority records, both of which supply a common data source to the Shared Shelf data entry environment. The schema then forms a mapping target for bulk data contributions, and also provides the template for individually contributed new work records. By linking the Data Warehouse to the data entry environment, contributors will be able to use existing BWR records or use Data Warehouse terms to normalize values in newly created Built Work records.

We have completed the schema for the full possible work record since some of the source collections (and future contributions) will have much richer data than the four required fields in the BWR schema; it would substantially lessen the strength of the registry to systematically "lop off" the rich data that will come with work records from Harvard, for example. The finalized schema has been developed with the BWR contributors (including the Getty Research Institute who have shared the schema design for their *Cultural Objects Name Authority).*

Data merging tools

With the full record (and our decision from year 1 about required fields from this superset), we have also made significant progress on programming that will match, sort, and perform preliminary matches for absorbing new datasets into the BWR. In year 2, we began a series of test protocols to identify the most efficient method for merging data from our five different seed data sets. The process for loading the five seed data sets (and subsequent institutional contributions) consists of loading each institution's contribution to a staging platform from which we have deployed mapping routines for testing the "new" contribution against the "core" set of records. We will be using the Harvard dataset (the largest and richest of the five seed sets) as the core against which to match other sets. But since the de-duplication and disambiguation routines are still in process, we have been conducting the test merges outside of the eventual software environment. Digging fairly deeply into the values in each dataset, we have seen their strengths and weakness. Later this quarter, we will have merged sample sets from all five source collections, that would serve as a demonstration of the results from the broader merge of such data.

The merging of the initial seed content has occurred outside of the Data Warehouse environment, but has been carried out with the eventual environment in mind. Each of these seed data sets was developed in separate and distinct data environments, and therefore need to be mapped to our new common data schema. As we compared sample sets, we discovered unsurprisingly a high percentage of duplication of built works both within and across seed sets. This prompted an ongoing process of de-duping and normalizing data in each set. The metadata and technical teams explored several approaches and are currently working to finalize workflows for data refinement and data merging. Some approaches include:

- Google Refine: While this toolkit makes the data refinement process somewhat easier, its maximum capacity for data 'crunching' is approximately 10, 000 records at a time. Therefore, an iterative approach was adopted for large sets such as the *Avery Index to Architectural Periodicals* which includes ~650,000 records. Even using this tool the process is labor-intensive and likely not scalable for many contributed data sets.

- GRI Vocabulary Program Data Loader and Processor: The GRI has developed automated tools for merging and refining contributed data for its other vocabulary projects (e.g. ULAN). While it still requires human intervention for final decision and authorizing of terms, the automated component is useful for preliminary matching and merging processes. The scripts for this toolkit were developed specifically for GRI programs, and it is non-trivial to adopt them to the ARTstor BWR infrastructure. But in order to leverage the toolkit for BWR, we are evaluating the efficiency of making a multi-step data merging workflow wherein, seed data sets prepared by the BWR team would be processed through the GRI toolkit, and then returned to BWR for ingestion into the registry for merging with the other sets, and for data enhancement (such as the geo-coding routine).

We have also experimented with what overlap and conflicts we will encounter when merging (again, at this point, outside of the eventual BWR algorithms) duplicate records. So, for example, all five seed record sets have individual records for Brunelleschi's *Ospedale degli Innocente* or Frank Lloyd Wright's *Fallingwater*. In some cases, the candidate record will have more data than the "core" record; in other cases, there will be conflicts between data elements and obviously the computer will have no means to deem which is the better data. We had originally hoped that the cell-by-cell addition of such conflicts could create a complete audit trail with all candidate additions or edits tracked (and eventually acted upon by community editors). At this point, the merging of all additions and conflicting data with source attribution on a field-by-field basis seems daunting. For the time being, we are looking at a model whereby the original contributor of a record would retain the capacity to edit that record, with candidate edits being managed in a wiki environment that would carry the whole alternative record and its source information as an appended field to the BWR core record. It would be available and someone with editorial privileges (either the core record contributor or other editors when we

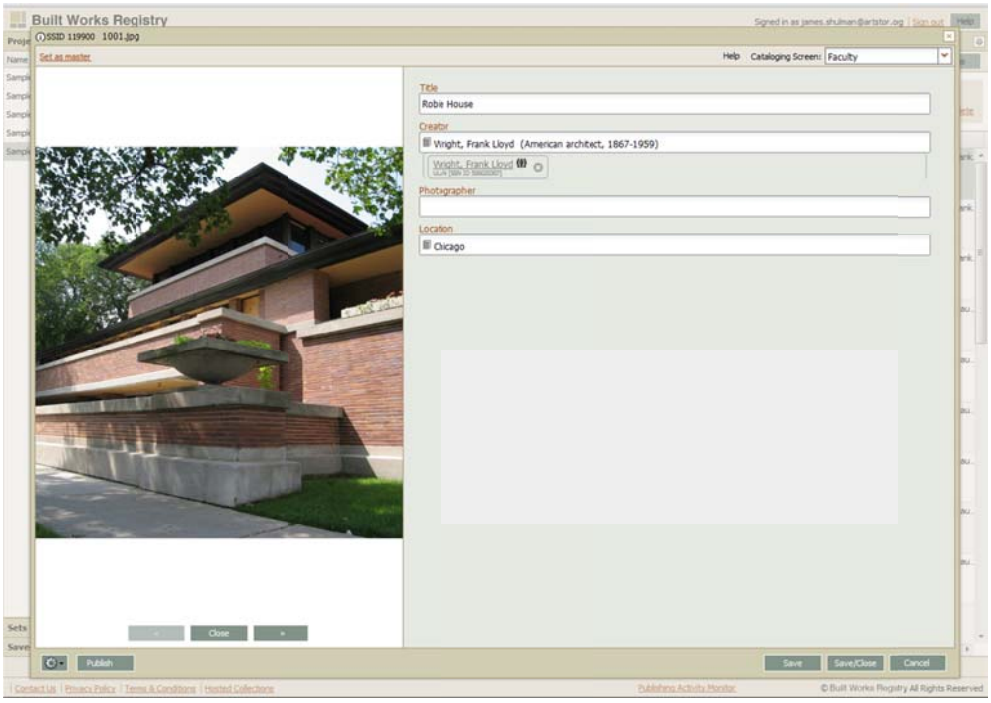
derive the requirements for such a role) could promote data elements from the appended candidate record to the core.

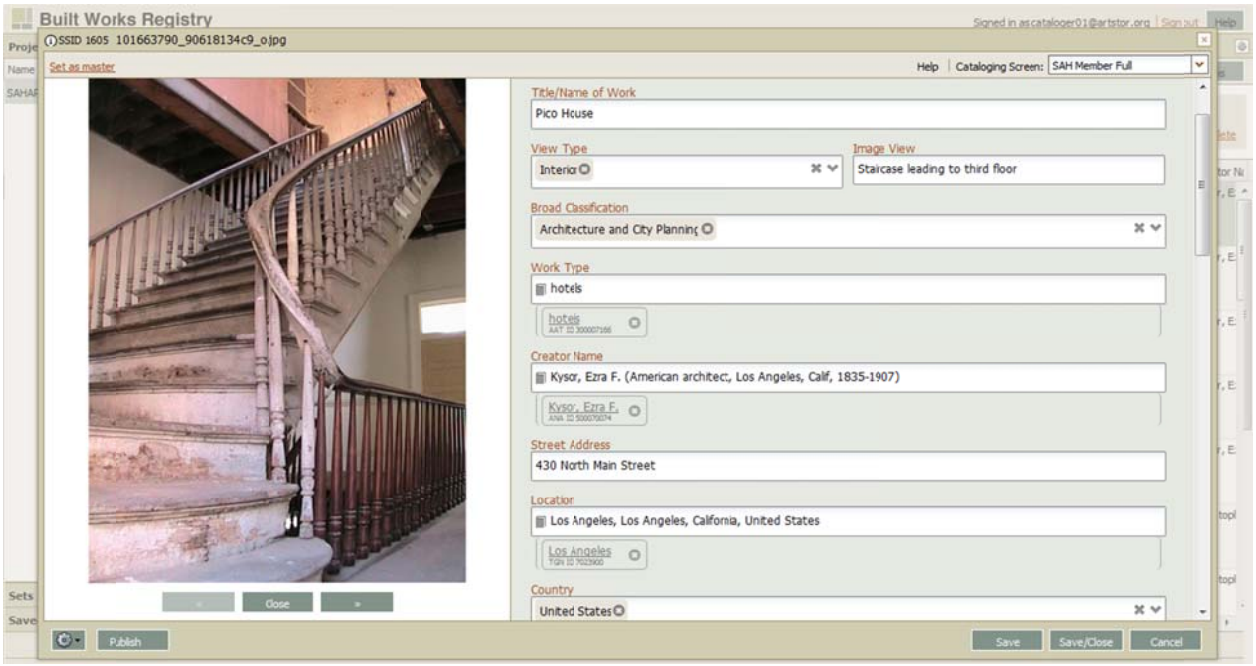
Record building and retrieving interfaces

If the BWR is going to grow over time, as we anticipate, it will need a robust environment that will facilitate the contribution of material and its use and since contribution will demand continual de-duplication and disambiguation (as described above), the tool (and data management environment) needs to be more powerful than the simple appearance might seem.

We envision some contributors to BWR will want to use very simple tools that allow them to fill in only three or four fields of cataloging information. This simplicity may make it possible for anyone who has an interest in buildings to use the system. At the same time, other contributors (primarily institutionally-housed catalogers), might be prepared to do complex cataloging where using authority controls for names, places, materials, and styles. Creating *each* of these two systems would present a different set of challenges, but we feel that it is important that BWR have both capacities. Obviously, with only four required fields, BWR is more likely to grow; but with the capacity to handle more full data records, the value of BWR to users of the service will be greater.

The record-creation screen will be a version of the Shared Shelf data entry screen, and will allow users to enter data into a simple set of fields, utilizing authority files where relevant. While the inclusion of thumbnail photographs (or larger images) is not required, it will be allowed as visual metadata to assist in disambiguation and the data can either be minimal (as with Screen Shot 1 below) or richer (as in the template shown in Screen Shot 2):





Geo-coding

An additional aspect of the record creation site will be tools for geo-locating the work since these data will provide the best chance of disambiguating works. This process is, unfortunately, not simple since one must SEARCH AND FIND the location before entering it. In other words, while a record creator can type in textual data ("Frank Lloyd Wright" or "Brick" or "Chicago" or "5757 S. Woodlawn Avenue") the linking of the record to geo-data (longitude and latitude) can only happen through a process of searching (by address or by visual navigation around a map) which then leads to a particular location (made manifest on a map interface) and then is highlighted (either by a point or by drawing the boundaries of the work) with a visual tool that selects areas on the map interface that link the underlying geo-data to the record.

In order to more fully explore current approaches, we convened an all-day meeting (July 11, 2012) bringing together experts and project directors from other institutions engaged in large-scale geo-location projects; these included:

- **Markus Flatscher**, Editorial & Technical Specialist**,** University of Virginia Press
  Project: SAH Archipedia http://sah-archipedia.org/

- **Sarah Williams**, Director, Spatial Information Design Lab, Graduate School of Architecture, Planning & Preservation (GSAPP) Columbia University
  Projects: http://www.spatialinformationdesignlab.org/

- **Matt Knutzen (**Assistant Chief, Map Division**), David Riordan (**Product Manager), and **Ben Vershbow** (Manager) New York Public Library Labs
  Projects: http://www.nypl.org/collections/labs

- **Chris Freeland (**Director, Center for Biodiversity Informatics) Missouri Botanical Garden
  Project**:** Biodiversity Heritage Library http://www.biodiversitylibrary.org/

We asked each participant to give a 'lightening talk' about their project(s) and to reflect on the merits and challenges each has faced in implementing a geo-coding component. Demonstrations and discussion surfaced a number of useful strategies and some encouraging suggestions for directions BWR might take, as well as highlighted what this expert group thought were the key high-value contributions BWR could make to all of our efforts:

- Assign unique IDs for each built work – *a BWR stated goal*
- Focus metadata development on more unique built works than the major known monuments;
- Refine and disambiguate records for iconic and lesser known works
- Leverage the scholarly community to contribute authority/disambiguation services to BWR
  - Create tools for community contribution
  - Integrate toolkit and problem sets into pedagogical setting
- Harvest large data sets from institutional contributor sources

For geo-coding in particular, participants proposed the following:

- Create an open-source web-services interface for generating geo-locations
- Allow for multiple and diverse geo-code 'statements' in each data record, i.e. create a 'geo-code data block' that will serve to cross-reference statements
- Enforce precision levels (implied hierarchy: rooftop, trace, XY latitude/longitude)

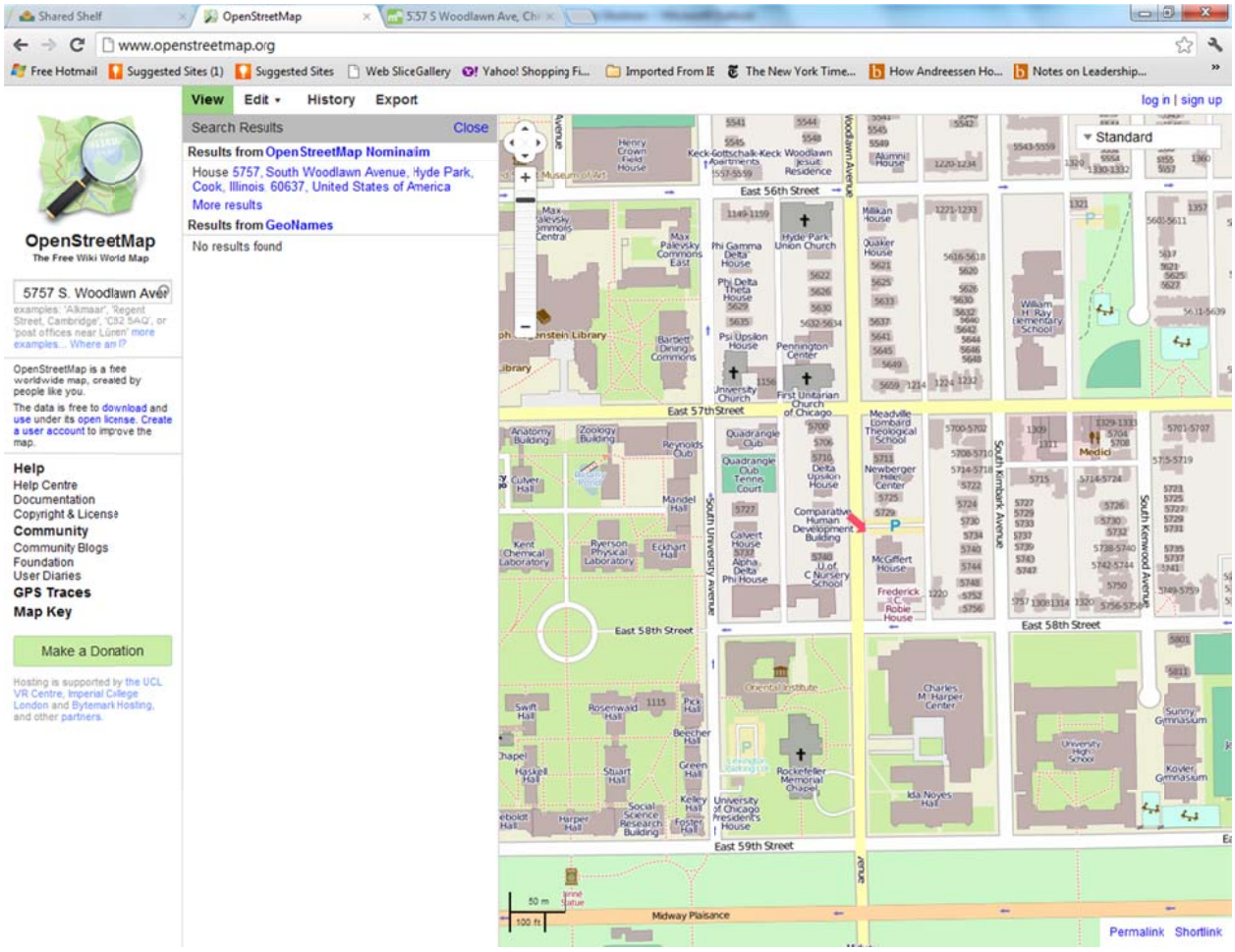
Based on the experiences of these colleagues working on other geo-coding projects, and on our own research into scalable strategies for BWR, we have decided to use a number of existing tools for various data layers that carry the least encumbering restrictions (unlike the popular and powerful Google Maps APIs which are encumbered by numerous legal obligations or restrictions).

We currently anticipate using a Mapquest API (but are also exploring the possibility of using the OpenLayers or Leaflet API) along with OpenStreetMap map tiles, supported by the search logic of Two Fishes.[1] The search logic layer enables data such as "84th and Park" to be translated into

[1] maps: http://developer.mapquest.com/web/products/open/map; http://www.openstreetmap.org/ https://github.com/foursquare/twofishes

map locations for the user to choose from. For the writing process of locating the work on the map tiles, we intend to use Leaflet (which will allow user to call out to an API that allows the user to pin or outline a site on the map. This stack of open geo-referencing layers is essential to the building of BWR records (since address location has limits and ambiguities) but it also will be essential to the retrieval of BWR records (see below).

Example of look up (of Wright's Robie House) using OpenStreetMap:



Read and Get site

BWR data will eventually be retrievable in two ways: (1) through a web-based discovery environment and (2) via a set of APIs that will search various fields in the repository without human intervention and allow BWR data to be employed in other software environments. Our current test site presents each work record (with a thumbnail image if the record has one) in

c. using the leaflet platform to build device independent geocoding UI http://leaflet.cloudmade.com/

response to textual queries, but we are only beginning work on enabling users to search via the geo-referencing stack that we are employing in data record creation. In the next year, we will need to determine how to do so, both for the look-up site and for the API access.

Addendum A:
ARTstor: new BWR project staff curriculum vita

Giovanni Garcia-Fenech
giovanni.garcia-fenech@artstor.org

Work Experience

2010 – Present: Communications Manager, ARTstor
Currently manages the editorial process for the creation of marketing materials in both digital and print formats, and promotes ARTstor and its resources, services, and projects. Creates content for the ARTstor Blog and oversees the company's social media strategies.

2007 – 2010: Communications Director, The Armory Show
Coordinated all aspects of press relations and advertising, and also contributed to the fair's marketing strategy. This included writing press releases, brochures, internal announcements, ads and website copy; proofreading, reviewing and editing all written materials issued by the company; and writing speeches and providing talking points for the director. Other responsibilities included developing and implementing public relations and advertising campaigns; drafting the annual marketing plan; supervising and coordinating ad design and the production of all printed materials and website maintenance; and organizing and coordinating press events.

2006 – 2007: Director, The Project Gallery
In addition to managing the production of exhibitions, I served as the liaison between the company and the press, museums and other galleries; I wrote press releases, spoke with collectors and visiting groups, and represented the gallery at art fairs and other public events. I also oversaw the design of advertisements and announcements.

2004 – 2005: Assistant Director, Roebling Hall Gallery
Responsibilities included overseeing the production of exhibitions and managing the gallery's inventory; being the point person between Roebling Hall and other galleries, the press and museums; writing press releases; speaking with collectors and visiting groups; representing the company at events.

2001 – 2003: Assistant Director, The Project Gallery
Wrote press releases; represented the company at art fairs and other events; managed the production of exhibitions, the inventory, and the shipment of incoming and outgoing work; oversaw a permanent staff of four and various freelancers.

1999 – 2001: News Editor, artnet.com
Edited a twice-weekly column for *Artnet Magazine*, an online arts magazine covering news, gossip, museum shows, awards, opportunities for artists, etc.; also wrote a monthly column reviewing the art scene in Brooklyn and contributed occasional articles on other topics.

1997 – 2005: Freelance Writer, various publications
Contributed news pieces about technological topics for wired.com; contributed articles about design for *Metropolis Magazine*; wrote articles and reviews for thing.net, fashiontribes.com and *Stitch Magazine*; wrote news articles for artforum.com.

Education
1993-95 School of Visual Arts, New York City Master of Fine Arts
1989-92 University of Houston, Texas Concentration: Painting
1991 Middlesex Polytechnic, Quicksilver Place, London, England Exchange studio program
1984-88 Austin College, Sherman, Texas Bachelor of Arts: Psychology, Art (double degree)

Addendum B: BWR Project Timeline

Trajectory	Activity	Project team effort	2010	2011				2012				2013		
			Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3
Project Admin	Hire/appoint/commit staff positions	Administrative team	X											
	Coordination, communications, documentation	Administrative team	X	X	X	X	X	X	X	X				
	Reporting	Administrative team		X		X		X		X				
	Dissemination	Administrative team	X		X		X		X					
Policies	Administrative Policies framework (governance, legal)	Administrative team	X	X	X	X	X							
	Content policies (scope, processes, workflow)	Administrative team			X	X	X	X	X	X				
	User policies (distributed curation & collaboration model)	Administrative team		X										
	Advisory group onsite meetings	Administrative team		X				X		X				
Metadata	Schema development	Metadata team		X	X	X	X							
	Data dictionary & guidelines	Metadata team			X	X	X	X						
	Processes / workflow	Metadata team			X	X	X	X	X	X				
Content	Identify/define	Metadata team	X	X	X	X	X	X	X	X				
	Extract & verify	Metadata team			X	X	X	X	X	X				
	Aggregate	Metadata team				X		X	X	X				
	Analyze	Metadata team			X				X	X				
	Normalize	Metadata team								X				
	Enhance	Metadata team								X				
	Migration prep to BWR data warehouse	Metadata team			X	X	X	X	X	X				
Technology														
	Policies design (processes, workflow)	Technology team		X	X	X								
	Technical design	Technology team		X	X	X								
	Backend infrastructure development	Technology team				X	X	X	X	X				
	Data warehouse development	Technology team				X	X	X	X	X	X			
	Export/ingest	Technology team												
	Deduplication	Technology team							X	X				
	User Interface development:													
	Data entry environment	Technology team					X	X						
	Administrative data management environment	Technology team												
	User management	Technology team												
	Shared infrastructure development:													
	Open interface	Technology team												
	Web services	Technology team												
	Integration	Technology team												
	Deployment	Technology team												
Evaluation		Admin, metadata & tech		X	X	X	X	X	X	X				