The charter is the foundational document that describes the rationale, goals, plan of work, resources needed, terms and conditions, and outcomes of a Center for Digital Humanities at Princeton (hereafter CDH) project. Charters are written by core members of a project team in a series of planning meetings taking place over the course of a month. The planning process is intensive, collaborative, and requires substantial input from everyone on a team. Charters serve as formalized agreements among all team members on such crucial questions as scope, technical design, infrastructural needs, and success criteria.

This is a digital copy of a “living document” at a single point in time. Charters are amended as necessary throughout the project lifecycle to document major changes and note when the “Built by CDH” Software Warranty and “Built by CDH” Long Term Service Agreement take effect, and serve as part of the CDH project archive. Certain components of the Charter, such as the Roadmap, are intended to be updated more frequently as development progresses and priorities naturally shift. For more information, see the Charter amendment policy located in the Agreements section of this Charter.

CDH charters and their planning documents exist in several forms as we have refined them over the years and tailored them to the several types of projects we have supported. For more about CDH project management, including the charter process, visit: https://cdh.princeton.edu/research/project-management.

Cite this document:

Marina Rustow, Rebecca Sutton Koeser,  Nicholas Budak, Gissoo Doroudian, and Rachel Richman, CDH Project Charter — Princeton Geniza Project Year 2, 2021-22. Center for Digital Humanities at Princeton. 2022. http://doi.org/10.5281/zenodo.7841822

# Princeton Geniza Project Charter

## 2021-2022 CDH Research Partnership — PGP version 4.0

*This charter goes into effect October 11, 2021 and ends October 11, 2022.*

# Overview

## Related documents

- For a list of terms and definitions used in this document and throughout the project context, see Terms and Definitions.
- For a list of common workflows used on this and other CDH sponsored projects, see Workflows.
- A list of crucial decisions made during this project's grant period and their rationale will be documented in the project Decision Log.
- 2020-2021 Charter

## Background

*For a detailed overview of project data changes during year one, see Appendix 1.*

The first year of the Princeton Geniza Project Research Partnership began in September 2020 amid the continuing COVID-19 pandemic, with all meetings taking place virtually. Prior to the establishment of the partnership, the Princeton Geniza Lab (PGL) had curated nearly 30,000 records, 4,500 transcriptions, and countless other supplementary materials describing Cairo Geniza documents since its establishment in 1985. The primary aims of the partnership in its first year were to reunite these silos of information in a single administrative interface and establish workflows that would ensure data integrity well into the project's digital future.

During the first year of the grant, the project's data progressed in accuracy, completeness, and structure. The team cleaned and reorganized document metadata in the main Google Sheet with assistance from CDH and DDSS, preparing it for import into a relational database. The team developed an ontology that distinguished physical "fragments" from documents in order to flexibly represent complex document types, including joins, that were difficult to handle in the existing system. CDH successfully migrated document metadata from Google Sheets into a relational database, preserving the edit history of document records while simultaneously merging and transforming them to fit the new schema. The HTR4PGP project was independently launched to investigate the feasibility of automated handwritten text recognition (HTR) for Geniza fragments, yielding a ground-truth corpus of 2,200 transcribed images.

Year one of the partnership began under the excellent project management of Stephanie Luescher (PhD student, Princeton, Near Eastern Studies). As the PGP had never had a PM before, Stephanie developed a new leadership model and expectations among the team that Rachel Richman (PhD student, Princeton, Near Eastern Studies) was thankful to inherit and continue to cultivate into the second half of the year. With changes as simple as inviting research assistants (RAs) to semi-regular meetings and taking advantage of digital tools such as Slack and Asana, Stephanie and Rachel helped create a team of PGP researchers informed about the broader goals of the project.

Alan Elbaum, an MD/PhD student at UCSF, was an invaluable senior researcher, consistently producing high level work both in service to the data migration as well as his own research goals (which frequently meant he was the first to spot bugs in the new system). He will continue in this role through at least May of 2022.

The more formalized structure of the PGP team, frequent meetings with the CDH, quarterly internal meetings, increased communication, and the use of project management tools all contributed to the successes we achieved in rolling out the new database. Morale is also high, with six out of eight of our RAs expressing interest in staying on board the project even as the school year resumes.

We achieved a majority of the project's in-scope outcomes for year one: we migrated document metadata to a relational database, created protocols to guide continuing work on describing documents in the database, designed a new web application that supports searching the entire PGP, and deployed a private "alpha" version of this application for use by the project team. Sustained collaboration between the PGP and CDH teams enabled the data migration to succeed, and the team has continued to reap the benefits of storing data in a relational database.

We did not achieve two in-scope outcomes for year one, and these are thus the major foci for the project's second grant year: migrating existing transcriptions to a new format that can be displayed as IIIF annotations, and the establishing new workflows for creating and editing transcriptions.

Our major goals for year two are to connect siloed data (transcriptions, metadata, images, links to related resources) in a single database for ease of management and to create and launch a public interface to make the data more amenable to research.

## Outcomes

### In scope

Outcomes in this section are the primary goals of the research partnership for this grant period (sorted roughly chronologically).

- Developing a public website with search and browse access to PGP database content with visual identity and standard content pages included on CDH projects (about, contributors, technical, how to cite, contact)
- Developing content management features for editing, publishing, and displaying content pages on the public site
- Migrating links for index cards and other attachments from the links database to PGP 4.0 database and making links editable by content editors
- Developing transcription workflow and deciding transcription format and location
- Creating a system for handling historical and ambiguous dates in a structured way to provide automatic conversion, sorting, and calculations
- Developing public search and browse interface for the contextual search feature that utilizes current tagging capabilities
- Creating workflows for publishing an initial dataset version and workflows for publishing periodic updates
- Writing team-authored dataset paper about the data publication
- Launching and promoting the new website

### Out of scope

Outcomes in this section are NOT slated to be worked on in this grant period, but CDH may consult on them as the Geniza team works on the data or planning for future work.

- Retiring and migrating the Index Card application at https://geniza.princeton.edu/indexcards/ into PGP 4.0
- Developing a database module, admin interface, design, frontend development for describing historical people related to PGP content
- Developing  a database module, admin interface, design, frontend development for historical places related to PGP content

## Significance

*Note: Statements for scholarly significance and design significance included in the 2020-2021 charter are still relevant and remain unchanged for the 2021-2022 charter.*

### Technical

#### Transcription data and workflow

Current solutions for creating and editing transcriptions for complex multi-lingual materials either require specialized training to use,  necessitate working switching between systems (e.g., eScriptorium) or are insufficiently structured (e.g., current web annotation implementations).  Regardless of the implementation we select for the new Princeton Geniza Project, our handling of transcriptions will of necessity be a novel addition to the current transcription and annotation implementation space. Our goal is to not only build a functional solution for the PGP for years to come, but also serve as a model or a solution for other projects.

### Historical and Uncertain Dates

The array of unknown, uncertain, and ambiguous dates needed to describe the Geniza materials accurately offer an opportunity to revisit and extend the partial date solution implemented for the Shakespeare and Company Project. We hope to adopt the Extended Date Time Format specification and make use of existing tools, most likely python-edtf, to meet the project team's needs for entering and working with dates. We also want to make it possible to use ambiguous dates for calculations, starting with sorting and filtering results on the public site, but eventually data analysis as well. PGP data and performance needs may well require contributing to and improving python-edtf, which would benefit a larger community of digital scholarship and digital curation work.

## Risks and dependencies

- Existing transcription data used in the current Princeton Geniza Project site is managed by Ben Johnston of the McGraw Center for Teaching and Learning. Migrating this data may benefit from consultation with Johnston.
- Transformed transcriptions in use for the HTR4PGP project, and newly generated HTR transcriptions, are part of the HTR4PGP project and managed in eScriptorium. Working with the HTR transcription imports will require some support from the eScriptorium team; timeline and availability of data will depend on that project.
- The index cards and attachments database are managed by Ben Johnston; migrating the attachment data and linking to the index cards application and other media files will require consultation with Johnston.
- Switching the new version of the PGP site live at the public url geniza.princeton.edu and will require coordination with Ben Johnston.
- Decommissioning the JTS Images site on the current http://geniza.princeton.edu/jtsviewer/ and including IIIF images for JTS content depends on PUL, and is dependent on getting the correct set of JTS images ingested into Figgy and published in a DPUL site.
- Transcription display, editing, and storage may require support from PUL, eScriptorium, or others depending on the chosen solution.

## Budget

The Princeton Geniza Lab will provide and manage compensation for all team members who are not CDH staff, including the Project Manager, Senior Research Assistant(s), and Launch Coordinator. PGL compensation will be administered  by the Geniza Lab Coordinator.

Costs for the project launch event and any swag will be shared equally between CDH and PGL.

# Roles and responsibilities

**Co-PI: Research Lead**
Marina Rustow

- Maintains and communicates overall vision for the project with Technical Lead
- Provides leadership on the research side of the project; oversees project's content aspects
- Communicates and resolves high-level project issues with Technical Lead
- Conducts acceptance testing if necessary to verify project features align with vision
- Attends regular project check-in meetings during development process
- Learns enough about project's technical components to discuss them publicly
- Supervises and maintains active communication with the Project Manager
- Notifies team in advance of any disruptions to work or unavailability
- Approves Project Manager's work on publicity, including project page on CDH site
- Credits CDH contributions in publications or talks about the PGP and keeps CDH apprised (as much as possible) about awards, recognition, or published research based on the new site's capabilities.
- Co-authors dataset essay with Technical Lead
- Recruits and helps manage launch coordinator

**Co-PI: Technical Lead**
Rebecca Sutton Koeser

- Maintains and communicates overall vision for the project with Research Lead
- Communicates and resolves high-level project issues with Research Lead
- Provides technical leadership; oversees design and implementation of project's technical aspects
- Adjusts choices of software tools and approach in order to achieve project goals
- Supervises software release process, including updating documentation
- Learns enough about project's scholarly aspects to discuss them publicly
- Can make project development decisions if Research Lead is unavailable

- Advises Project Manager on design and modification of project workflows
- Advises Project Manager on creation and modification of project data protocols
- Contributes to development, documentation, automated testing, and code review of software for the project
- Coordinates development and design work with other CDH projects and priorities
- Supplies periodic updates on project status and progress via PM tooling (e.g. Asana)
- Communicates and coordinates with technical collaborators and consultants on campus and elsewhere (including Ben Johnston at McGraw)
- Co-authors dataset essay with Research Lead

**Project Manager (PM)**
Rachel Richman

- Schedules CDH project check-in meetings
- Schedules meetings with CDH and project collaborators (DDSSI, others)
- Creates agendas for project meetings; assigns facilitator(s)
- Arranges meeting note-taking in coordination with Research Lead
- Identifies meeting action items and creates tasks using PM software
- Maintains awareness of current dev/design team work during active development
- Maintains and, if necessary, creates workflows to help team achieve its goals
- Conducts acceptance testing of project features, escalating to Director if needed
- Oversees day-to-day work of project data team for tasks specific to CDH-Geniza partnership
- Assists Research Lead in managing launch coordinator
- Documents or assigns someone to document decisions made in meetings or on Slack in the project Decision Log
- Facilitates project retrospectives

**User Experience (UX) Designer**
Gissoo Doroudian

- Chooses appropriate design tools for the project
- Conducts design testing of project user interface (UI)
- Makes decisions regarding project user interaction flow, UI: layout, color, and typeface, and visual design incorporating team input at design meetings
- Initiates collaborative and iterative design for information architecture (IA) and UI
- Incorporates usability and accessibility principles in designs
- Conducts UX research and/or usability testing when crucial prior to making design decisions

**Research Software Developer**
Kevin McElwee

- Develops and documents software for the project
- Writes automated tests for software for the project
- Assists Technical Lead in choosing software tools and approach to the project
- Reviews code contributions from other Developers
- Delivers software features for acceptance testing; writes testing notes/instructions
- Delivers UI components for design testing; writes testing notes/instructions
- Assists Technical Lead in consulting and advising project team on data questions, workflows and tools
- Assists with preliminary analysis and exploration of datasets

**Senior Research Assistant (SRA)**
Alan Elbaum *(MS, UC Berkeley; MD candidate, UC San Francisco)*

- Creates and maintains project data protocols with Research Lead and Project Manager, in consultation with Technical Lead, Research Software Developers, and UX Designer
- Maintains broad knowledge of scope and shape of project data via frequent interactions directly with the data

**Research Partnership Advisor**
Natalia Ermolaev
Meredith Martin

- Advises Development and Design Team on high-level project direction (i.e. charter review, amendment review)
- Communicates overall CDH strategic and research priorities to Research Lead
- Facilitates collaboration between units (in this instance, McGraw Center for Teaching and Learning and CDH)

**Project Management Mentor**
Natalia Ermolaev

- Trains and mentors Project Manager in essential project management skills
- Helps troubleshoot project or team-related issues

**Administration and Finance**
Deena Abdel-Latif *(Geniza Lab Coordinator)*

- Manages project funds
- Tracks and reports PGL team hourly work
- May assist PM with scheduling meetings
- Assists with coordinating launch
- Assists with ordering and distributing launch swag

**Launch Coordinator**
Undergraduate assistant (TBD)

- Helps to coordinate project launch
- Coordinate and publish any publicity materials related to project launch (press release, social media, etc.) with approval of PIs
- Help coordinate purchasing and distributing launch swag

**Communications Specialist**
Camey VanSant (*CDH Communications Lead*)

- Leads launch strategy and planning (setting timelines, assignments, etc)
- Works with PIs and Launch Coordinator to plan the launch
- Assists with project publicity, including project page on CDH website, with assistance from Technical and Research Leads as needed

# Roadmap

The roadmap is a high-level, strategic overview of planned development and design work for this grant period. It focuses on take-aways and decisions rather than products and is intended to be modified as the project progresses. Near-term events will necessarily have more detail than longer-term ones, reflecting the inherent uncertainty of digital product development.

Prior to the end of each of the phases listed below, the Project Manager will facilitate a retrospective meeting at which team members will share and reflect on progress and lessons learned. At this meeting, the team will collaboratively outline a short statement that encompasses the major take-aways from the retrospective. The Project Manager will assign team members as needed to fill in details in the retrospective summary statement.

Prior to starting development at the beginning of each of the phases below, the Technical Lead will propose and circulate a more detailed roadmap for the next phase, based on the retrospective and current project priorities. The updated roadmap will be circulated to the project team for discussion and must be agreed upon by Technical Lead and Research Lead.

## Fall 2021 (October–early February)

The primary goal for the fall is a minimal version of the new public site that can be made live so the old site can be retired. See Appendix 2. Features for Minimum Viable Product (MVP) for the list of agreed upon functionality.

## Initial public site (MVP — PGP 4.0)

- Design remaining features required for MVP
  - Logo, visuals, colors
  - Site navigation
  - Home page & content page styles
- Develop remaining features required for MVP
  - Refine and extend existing public search implementation and detail pages
  - Import index cards / attachments metadata into database and make editable
  - Preliminary implementation of transcription display
  - Implement typography, visuals, and styles per accepted designs
  - Preliminary content management for site pages

## Refine date handling

- Review data work and text extraction for historical and converted dates; compare existing converted dates in current data with conversions generated by the python convertdate library
- Explore and propose solutions for automatic date conversion
- Explore and propose solutions for sorting on and calculating with ambiguous dates basing requirements on examples from current data

## Transcriptions requirements

- Gather and document requirements for transcription display, edit, and management functionality
- Prototype and pitch possible implementations (TEI, IIIF, METS/ALTO)

## Contextual Search

- Develop tagging guidelines
  - Develop a logic for eliminating and revising tags
- Develop dynamic tag co-occurrence network visualization as reference for data and design work on tags and contextual search
- Experiment and make a decision on the idea of “regions” for a group of closely associated tags
- Explore and experiment with directions for visual design using the tags

## Planning

- Recruit undergraduate assistant to help coordinate the launch
- Begin planning launch

## Deliverables

- Initial public site live at geniza.princeton.edu
- Proposed solution for date handling

- Requirements document for transcriptions

## Winter/spring 2022 (mid February-May)

### Transcriptions

*Goal:* Determine storage format for transcriptions and migrate existing content from TEI. Support adding new transcriptions (i.e., existing transcriptions stuck in Google Docs) with block-level alignment.

*Transcription MVP:*

- *Determine and document IIIF Annotation format for transcriptions*
- *CRUD for transcriptions (create, read, update, delete)*
- *Convert existing TEI content to IIIF Annotation*
- *Support adding new block-level transcription text*
- *Out of scope:*
    - *Bulk import, transcription review, line-by-line*

- Implement transcription editing solution based on IIIF annotation with Annotorious, using PUL's Simple Annotation Server as storage backend
    - Prototype specific portions of the interface and editing functionality to confirm approach and get feedback
        - wysiwyg editor, image selection tools, line-by-line layout
        - Test importing and editing eScriptorium content
    - Implement storage plugin for Annotorious + annotation server
    - Integrate wysiwyg editor for Annotorious
    - Implement script to synchronize transcriptions to a GitHub repository for backup, versioning, corpus dataset
    - Migrate existing TEI to annotation list format and import into annotation server
        - Pre-step: cleanup wrapped lines without line numbers in TEI that were preserved from layout in printed transcriptions
        - Update search indexing to work with new format and storage (both on-demand indexing and bulk)
        - Update display to work with new format and storage
        - Link transcription content to appropriate scholarship record (replaces current sync TEI script)
    - Embed the same public image/text display and edit functionality in the admin document edit view
- Refine image/transcription display with alternate views, layouts, interactivity
    - Translate public site requirements into design requirements
    - Design display and interactions
    - Implement interactions

- Improve IIIF image availability & integration
    - Generate local manifests to segment where source manifest does not match our shelfmark segmentation
    - Generate local static iiif manifests and iiif zero for content where images are available and permissions allow (i.e, Bodleian)
    - Update code to support iiif presentation api v3 and iiif image api level zero

## Refine and improve public site

*Note: improvements to the public site need to be prioritized by Research Lead, since we may not get to all of them.*

- Logo and image work
- Usability testing
- Implement additional designs and functionality as prioritized by Research Lead and Project Manager in conversation with Technical Lead. Chosen from planned functionality and existing designs. May include:
    - Display images on search results
    - Additional search filters
        - Has transcription
        - Has translation
        - Has discussion
        - Has images
        - Document date range filter *(dependent on/included in date implementation)*
        - Designed filters include a location filter, but we don't have data for this; should we consider language+script filters?
    - Additional search sort options
        - Random *(new)*
        - Date *(dependent on/included in date implementation)*
        - Input date
    - Refine document details page designs and implementation
        - Display primary and secondary languages
        - Link to documents that appear on the same fragment
        - Revise designs to reflect implementation changes and incorporate all available data from the backend
            - Design related documents tab (replaces External links in earlier designs/sitemaps)
            - Improve linking to external sources
            - Recommend looking at FPG if we don't have an image
            - Implement document detail page changes
            - Implement related documents tab
    - Tuning the search
        - Language-specific indexing for transcription text where possible
        - Further customize field boosting

    - Enable conversion of Arabic search terms so they match and highlight Judeo-Arabic content
        - How to read Judaeo-Arabic manuscripts 5.1_Dec2021
    - Social media preview image when sharing links
        - Design a default social media image based on logo and site visuals
        - Fragment image preview when available

## Document Date implementation

*Document date MVP:*

- *Automatic conversion in admin for supported calendars*
    - *Add support for additional calendars iteratively*
- *Display document date on document details page (original and converted)*
- *Implement date range filter in document search for known, converted dates*
- *Implement sort by date in document search for known, converted dates*

- Investigate dates entered and test automatic conversion
    - Review data work and text extraction for historical and converted dates; compare existing converted dates in current data with conversions generated by the python convertdate library
- Implement automatic date conversion in admin interface
    - handle simple cases first; add support for more complex cases iteratively
- Determine approach for sorting on ambiguous dates and date ranges
    - Implement conversion and index date and date range in solr
- Incorporate dates into public interface
    - Display in search results and detail page
    - Sort search results by document date
    - Filter results by document date

## Contextual Search (exploration and design)

- Explore options for generating clusters (Vineet)
    - Keyword assisted topic models
    - Word2vec & affinity propagation
- Initial design of the search/browse interface
- Conduct usability testing with appropriate stakeholders

## Hebrew/Arabic version of the site (preliminary)

- RTL interface
    - Design RTL components
    - Hebrew fonts and typography
    - Implement RTL components for public site
    - Implement Hebrew typography
    - Coordinate Hebrew translation for site components and page content
- Translation support

    - Implement middleware to differentiate between supported translation languages and public site languages
    - Make document type and document descriptions translatable in admin
    - Display translated document types in search and document details
    - Index and search on hebrew description when available; fallback to english description as needed

### Planning

- Plan for launch
- Plan project wrap-up. Identify remaining needs and priorities, determine handoff.
- Discuss future of the project and any long-term plans; identify grant opportunities
    - Support application for Research Computing RSE

## Summer 2022 (June–August)

### Transcription improvements

- Import eScriptorium content
- Implement line-by-line editing interface
- Auto-generate zones for transcribing text
- Support transcriptions for documents without images
- Workflow to convert block-level transcriptions to line-level

### Bulk record import

- Define required and optional fields and share a spreadsheet template
- Adapt relevant portions of original import code to support bulk import
    - If pgpid already exists, or shelfmark matches an existing record, append description to internal notes field and flag as needing review

### Additional refinements and improvements to public site

- Continue to improve public site functionality and designs as needed and prioritized by Research Lead and Project Manager in conversation with Technical Lead, incorporating input from UX Designer based on user feedback and usability testing
    - Switch transcription display from Judaeo-Arabic to Arabic
- Improve transcription workflows as needed
- Improve date handling as needed
    - Add separate fields in admin for dates ranges and uncertain dates
    - Display on document details page
    - (unclear if these can be used for search/sort)

### Data exports and publication

- Determine dataset publication platform/approach
    - Create a regular, automatic export that syncs to a GitHub repository and makes the data available via Flat GitHub

   - Published versions based on tagged releases from this repository
   - Setup regular data validation to make publication easy
- Create preliminary public dataset export (metadata)
  - Start with fields in current admin data export and iterate from there
    - Make available as both CSV and JSON
  - Determine how to integrate with transcription data; likely two separate datasets; use PGPIDs to link
  - Implement regular dataset publication workflow; publish initial version
  - Optional: dynamic, filtered export from the public search interface for smaller sets (limit?)
- Outline and begin drafting dataset paper (Co-PIs)
  - Review example essays that could serve as models (Shakespeare and Company Project datasets essay recently published with *Cultural Analytics*, or other *CA* dataset essays; Ben Lee's "data archeology" of his Newspaper Navigator dataset published with DHQ)
  - Consider possible venue(s) for publication and audiences
  - Draft an outline of what should be included and who will be responsible for which sections
  - Preliminary analysis and exploration of datasets (co-PIs); identify data analysis and data visualizations to include
  - Draft sections about the history of PGP (Research Lead)

### Contextual Search

- Design the finalized search/browse interface
- Integrate clustering into the PGP application based on Spring exploration
  - Display and navigate clusters based on design
  - Toggle between regular search and clusters
  - Display cluster membership on document detail page
- Cluster management
  - Script to regenerate clusters periodically
  - Admin interface to review regenerated clusters
  - Logic to fit new documents to existing clusters between when they are regenerated

### Hebrew/Arabic version of the site (public)

- Design Arabic and typography
  - Implement Arabic typography
  - Coordinate Arabic translation for site components and page content

## Fall 2022 (September–early October)

### Wrap up

- Continue drafting dataset essay and data analysis/visualization
- Complete any in-progress development and design work; address any high priority bugs

- Identify opportunities for additional publications, presentations, and classroom integration for the work
- Discuss and plan next steps for the project

The CDH Long Term Support Agreement (see 2018 version) will take effect when the charter year ends, unless the co-PIs determine otherwise. The CDH Warranty will not apply to this project, since the initial public launch will happen early in the project year with plenty of time to address problems while the project is in active development.

# Agreements

## Charter amendments

While the Roadmap section of this document is expected to change as the grant period progresses, the Roles and Responsibilities and Agreements sections represent mutually binding terms between the CDH and its research partners. Adjustments to either of these two sections require an amendment document describing the proposed changes to be submitted for review by the CDH Research Partnership Advisor. Amendments are subsequently signed by the Research Lead and CDH Research Partnership Advisor to incorporate them into the project's charter.

## Wrap-up

At the conclusion of the grant, the Project Manager and the Technical Lead will collaborate on a short (less than 500 words) summary of work that was accomplished during the grant period. The summary should describe the goals reached and outcomes of the project, and explain major changes and discrepancies with planned work. Continuing projects will include this information in the charter overview section for the subsequent phase of the project.

## Project pause

To ensure that all projects receive sufficient and equitable staff time, time-sensitive development and design queries and requests must be addressed within two weeks of initial (email) request. The Research Lead is responsible for communication with the CDH Development & Design team. If the Director does not respond to a task that has been indicated as time-sensitive by the CDH team within 2 weeks of initial request, further project development will be paused until the project can be reasonably integrated back into the CDH development schedule.

## Rights, permissions, and attribution

Site content and data will both be licensed under Creative Commons Attribution 4.0 International  (CC-BY 4.0). If any of the datasets consist solely of factual data where authorship cannot be claimed, they will be licensed as CC0.

Any software developed by CDH that merits release will be licensed under Apache 2.0. The Technical Lead will fill out an invention disclosure form in order to gain approval from the Office of Technology Licensing in order to release the code. Before approval is granted, the code will be owned by the Trustees of Princeton University.

## Credit

All team members will be credited on the project's website and CDH project page. The project's website will include a sponsorship statement (indicating the CDH as well as any other supporting groups, departments, agencies) and will include a citation statement indicating how the project assets should be cited. The site will also list and link to other projects that contributed data. All contributors will be included in citations for the project as a whole or individual assets, including datasets.

## Labor

Academic labor inherently involves imbalances of power. In accordance with CDH values, team members commit to ensuring that all project labor for this grant year will be compensated and credited — particularly with respect to contributions by undergraduate students, graduate students, and other vulnerable groups. Issues with compensation should be forwarded to the CDH Business Manager.

## Future Grants Applications and Software Development

The Research Lead will give CDH ample lead time to review any future grant applications that involve building on, updating, or expanding CDH work on the project.

The Research Lead will include the Technical Lead in conversations about hiring other groups to contribute software development to the project as long as CDH is hosting and maintaining the project.

# Appendix 1. Year One Project Data Overview

## Transcriptions

- Researched and found the sources or editors for transcriptions from pre-2015 whose source or editor hadn't been (or was no longer) noted.
- Goitein scans:
  - Generated a list of documents that had no transcription in PGP but did have a Goitein typed transcription attached; **these remain to be input as transcriptions.**
  - Generated a list of Goitein scans that weren't attached to any record; attached some of them to records. **Others remain to be attached.**
- Finished correcting OCR'ed transcriptions of ~700 documents from Gil, *In the Kingdom of Ishmael* (1997).
- Reviewed scholarship to pick up transcriptions that had been inadvertently omitted from volumes of transcriptions already included in the legacy PGP data (such as Gil, *Pious Foundations*).
- Added hundreds of new transcriptions, both from scratch and from unpublished dissertations.
- Began generating a list of books and articles with transcriptions that had never been entered into PGP.
- Established a consistent set of transcription conventions for the first time in PGP's history.

## Descriptive metadata

- Rewrote dozens of convoluted document descriptions and wrote hundreds of "awaiting description" descriptions.
- Eliminated the "technical notes" and "notes2" fields and parsed out their data to specific fields.
- Provided "input by" and "input date" information for thousands of records and cleaned the "input" fields for thousands of others.
- Rewrote hundreds of stub descriptions previously labeled as "documentary according to FGP" (i.e., FGP's crowdsourced or automatically generated descriptions of documents, all of which needed checking by an expert).
- Wrote hundreds of new descriptions previously labeled "see Goitein's notecard."
- Cleaned up and restructured hundreds of bibliographic records.
- Found missing bibliographic data for hundreds of records (e.g., Hebrew article titles).
- Added descriptions for 19th-20th c geniza fragments from the Basatin cemetery that are catalogued in Arabic.

## Data structure

- Established our new fragment vs. document ontology.
- Applied new ontology to complex documents (e.g., more than one fragment to a document—joins; multi-folio texts—and more than one document on a fragment).
    - Disaggregated hundreds of records that included more than one document on a fragment (e.g., recto and verso records in a single PGPID) and moved their attachments.
    - Prepared for upcoming merge of many duplicate records.
- Restructured the libraries and collections fields.
- Established language/script ontology.
- Added languages tags to hundreds of records.

## Images

- Started expanding our image coverage by gathering spreadsheets correlating shelfmarks and IIIF links from the Bodleian, the British Library and Penn.

## Related work ("bonus projects")

- Developed a database for currencies in the Geniza to help scholars date documents by currency references.
- Mined all of the dated documents to create charts cross-listing individuals and their dates.
- HTR:
    - Established core corpus of 2,200 images for which PGP has transcriptions for use as ground truth (GT).
    - Divided core corpus into simple and complex layouts.
    - Automatically segmented and manually corrected segmentation for simple layout documents.
    - Established transcription conventions (see above).
    - Wrote code to automate the importation of PGP transcriptions into e-Scriptorium and the stripping out of editorial annotations (stuff in brackets etc).
    - Began manual correction of PGP transcriptions of simple layout fragments.

# Appendix 2. Features for PGP Site version 4.0 Minimum Viable Product (MVP)

Agreed upon functionality required for initial public release of new PGP interface.

*Note: this is a smaller subset of the features that are in scope for this charter year, since we are targeting an early public version as soon as possible.*

- Search documents
    - Keyword search across shelfmark description, transcription, tags, languages, and other fields currently indexed
    - Filter by document types
    - Search results with title, input date, PGPID, description excerpt, scholarship indication
    - Keywords in context for description
    - Keywords in context for transcription
    - Configure relevance boosting for shelfmark field
    - Pagination for search results
    - Sort by relevance
    - Sort by scholarship records (most/least)
    - **NOT INCLUDED:** features in accepted search designs, including additional search filters, sort options, images in search results, document dates; further refining and configuring search boosting; cluster-based browse
- View details for individual documents
    - Document details per accepted designs, including
        - Editor, input date, permalink, description
        - Image viewer if IIIF images are associated
        - Transcription text for transcriptions currently included in PGP (i.e. TEI transcriptions)
        - Transcription line numbers
    - Scholarship records
        - Revised design for a more bibliographic layout
        - Revise implementation based on revised design
    - Basic display of index cards with link to current index cards site
        - Basic management of index cards in site database
    - **NOT INCLUDED:** external links page for individual documents
- Home page and basic content pages
    - Implement styles and visuals for home page, content pages (about, contributors, technical, how to cite, contact), site navigation

    - Wagtail configuration for basic site content editing and management
- Visual design: logo, typography, color