Are digital humanities projects sustainable?
A proposed service model for a DH infrastructure

CNI MEMBERSHIP MEETING: FALL 2018
MONDAY 10 DECEMBER 2018 2:30-3:15PM
CHRISTINE MADSEN & MEGAN HURST
ATHENAEUM21
The Problem

A proliferation of DH projects, tucked away in more than 18 departments

Total Count = more than 200
What is the “minimum viable service” for a digital humanities infrastructure that would be used by a maximum number of digital humanities researchers?

In other words, what is the minimum functionality required to persuade researchers to use a centralized, supported, and sustainable digital infrastructure, rather than create something themselves, or use commercially-available tools.
The Research: Interviews + User Needs Analysis

- Interviewed 31 people from the Humanities and Social Sciences, representing 25 projects
- Reviewed all their available projects for documented user experience and user needs
For ~40 projects we:

- Approached each online project as an end-user
- Verified the functional requirements
- Double-checked the proposed “minimum viable service” against each project
The Findings: 4 Areas

1. What do DH researchers have? What are their research data?
2. What do people want to do with the data they have?
3. What are the functional requirements for sustaining these projects?
4. What are some of the functional solutions?
What do people in DH study? What are their research data?

In order of frequency:

1. Metadata (descriptions of things)
2. Text (full, transcribed text of things)
3. Images
4. Audio
5. Video
6. Software (but very little)
What do people in DH study?
What are their *research data*?

In order of frequency:

1. Metadata (descriptions of things)
2. Text (full, transcribed text of things)
3. Images
4. Audio
5. Video
6. Software (but very little)

Good news! This is largely not a software preservation problem!
Findings

1. There is a limited number of research data types
What do people *mostly* want to do with their research data?

1. **Search and find**
2. **‘Publish’ online** (make available in a browser, via a stable, permanent URL)
3. **Compare versions**
4. **Download**
5. **Listen / watch**
6. **Transcribe**
7. **Analyze**
8. **Run software**
What do people *increasingly* want to do with their research data?

1. Search and find
2. ‘Publish’ online (make available in a browser, via a stable, permanent URL)
3. Compare versions
4. Download
5. Listen / watch
6. Transcribe
7. Analyze
8. Run software
   - Map
   - Visualize
   - Machine learning
   - Visual search
Findings

1. There is a limited number of research data types
2. There is a limited number of required functionalities
So, what is the problem?

What are the challenges associated with DH projects?
First things first

WHAT DO WE MEAN WHEN WE SAY SUSTAINABILITY?
archive (noun) - 1. A collection of historical documents or records providing information about a place, institution, or group of people.
1.2 A complete record of the data in part or all of a computer system, stored on an infrequently used medium.
archive (verb) - 1. To place or store (something) in an archive.
1.1 Computing Transfer (data) to a less frequently used storage medium such as magnetic tape.
Glossary: What are the issues here?

**archive (noun)** - 1. A collection of historical documents or records providing information about a place, institution, or group of people.  
1.2 A complete record of the data in part or all of a computer system, stored on an infrequently used medium.

**archive (verb)** - 1. To place or store (something) in an archive.  
1.1 Computing Transfer (data) to a less frequently used storage medium such as magnetic tape.

**preservation (noun)** - The **action** of preserving something.
<table>
<thead>
<tr>
<th><strong>archive (noun)</strong> - 1. A collection of historical documents or records providing information about a place, institution, or group of people. 1.2 A complete record of the data in part or all of a computer system, stored on an infrequently used medium.</th>
<th><strong>archive (verb)</strong> - 1. To place or store (something) in an archive. 1.1 Computing Transfer (data) to a less frequently used storage medium such as magnetic tape.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>preservation (noun)</strong> - The action of preserving something.</td>
<td><strong>sustainability (noun)</strong> – 1. The ability to be maintained at a certain rate or level. <strong>sustain (verb)</strong> – 3. Cause to continue for an extended period or without interruption.</td>
</tr>
<tr>
<td><strong>PASSIVE</strong></td>
<td><strong>ACTIVE</strong></td>
</tr>
</tbody>
</table>
Can a data repository be the answer for sustainability?

No. repositories are

- ...archives
- “...is not for the storage of data that is still in use by research projects.”
- ...requires ‘packaging’ the data in a way that prevents granular access

Sustainability requires access without interruption.
- Maintaining a level of access to the data intended by the researcher

It is a good idea to archive the data from these projects, but that will not sustain them.
Requirements for Sustainability

Sustainability requires understanding at least three things:

- What is essential to sustain
- What should not – or need not be – sustained
- What is unique about these projects?
What is Unique About these Projects?

- Bringing together a collection and/or a corpus for the first time
- Providing new forms of access to that content by making it electronic and searchable

To be clear:
- The content / collections / corpora are not usually unique
- The software is not usually unique

But
- The methods of access provide the opportunity for new scholarly opportunities
DH Workflows: A Deep Dive

THERE ARE MORE WORKFLOWS THAN WE THINK
“Traditional” Research Data Workflow

Collect / Create

Archive / Preserve

Organize

Publish

Analyze
The Reality of the Data Workflow
The Reality of the Data Workflow

This process could take 10-100 years

Collect / Create
Organize
Update
Publish Results
Make Available
Analyze
Analyze
Publish Results
The Reality of Data Lifecycles in DH

- The ‘research data’ being created is not just data, it is corpora, collections, and reference works.
- Think of it more like a dictionary than ‘traditional’ research data
  - Aggregations of granular data
  - Long-term activity
  - Data is ‘shared’ and made public much earlier in the workflow than in the traditional workflow diagrams
  - Multiple research projects using the data at the same time in different ways
  - New research leads to corrections, additions, and updates to the data (as well as ‘publications’)
- Not unique to DH – think Human Genome project or longitudinal, multi-generational medical studies
An Analogy: The Oxford English Dictionary
Discover the story of English
More than 600,000 words, over a thousand years

Welcome to OED Online. If you or your library subscribes, dive straight into the riches of the English language. If not, click on the images below to learn more about the OED, see What's new, or take a look at Aspects of English, our language feature section.

Hobby words
Help us to identify and record the words, phrases, and expressions particular to your hobby or pastime.

Share your words

Word of the day
† side-glass, v.
1679
transitive. To gaze at (a person) amorously or flirtatiously through...

Sign up for word of the day

Recently published
entrammel, v.
You don’t archive the OED when you are ‘done’, you expose it for research and analysis. That is how you sustain it.
Oxford Dictionaries API

Enhance your app with our world-renowned dictionary data.

GET YOUR API KEY
This is not how the OED Works
This is how the OED works

Collect/Create
Organize
Update
Publish Results
Analyze
Make Available

This has taken 100+ years
First fascicle released in 1884
Changing the language

Rather than talk about ‘research data’ we should talk about DH projects as producing corpora and reference collections
Findings

1. There is a limited number of research data types
2. There is a limited number of required functionalities
3. **Sustainability requires sustained, granular access**
   
   E.g. ‘maintained at a certain rate or level’ (from the definition)
What Each Project Needs: Infrastructure

- A way to create metadata (that is, to describe things)
- A place to put ‘data’ (text, images, video, audio)
- An index that allows end-users to search and find things
- Ways to render these objects in a browser with stable/permanent URLs so they can be cited
- A place to engage and innovate – that is, to do more experimental things like image matching, visualization, etc.
- A way to update the data
What Each Project Needs: People

- People to help translate functional requirements into technical requirements
- People to maintain, manage, update the software and storage
- Expertise in hardware, software, data and metadata standards
- People to sustain the collections and data and to migrate formats when needed
- Support for fundraising
- Expertise in outreach
What is needed to sustain these projects in aggregate?

1. People
2. Storage
3. Software
4. People
What is needed to sustain these projects in aggregate?

1. People
2. Storage
3. Software
4. People

**People** to help ‘translate’ functional requirements into technical requirements

**People** to maintain and update the software
What is needed to sustain these projects in aggregate?

1. People
2. Storage
3. Software

Infrastructure that allows continued (long-term), item-level access to these collections and corpora. (Also includes people to help manage/preserve)
1. There is a limited number of research data types
2. There is a limited number of required functionalities
3. Sustainability requires sustained, granular access
4. Sustainability requires a mix of technology and people
Findings

1. There is a limited number of research data types
2. There is a limited number of required functionalities
3. Sustainability requires sustained, granular access
4. Sustainability requires a mix of technology and people
5. There is no, single, out-of-the-box solution to meet all these needs
Option 1: Provide Storage and People for Each Project

1. Give projects storage
2. Hire a team of people to look after them

**Pros**
- Each project has full autonomy
- Funders like to give money for something ‘new’

**Cons**
- Not scalable
- Who hires/manages the people?
- Doesn’t solve the long-term problem because eventually people will no longer have funding or project knowledge – then what?
Option 2: Provide Sustainable ‘Service Layers’

- Collect
- Create
- Organize

- Project 1
- Project 2
- Project 3
- Project 4

- Storage and Preservation
- Access / Discover
- Basic Access
- Engagement Analysis or Innovation

- Collect
- Create
- Organize

- Text
- Images
- Audio / Video
- Data
- Metadata

- Aggregated access
- APIs / SOLR Index
- Individual Item Access
- Image recognition
- Video/ audio player
- Digital Edition Viewer

- Lightweight software for object-level discovery and access
- Present (Project websites)

- Present

- Engaged
- Analysis or
- Innovation

- Life of the funded project

- Lifespan “forever”
- Lifespan 5-10 years
- Lifespan 3-5 years

- Life of the funded project
Storage and Preservation

- Text
- Images
- Audio / Video
- Data
- Metadata

Storage / Preservation Layer
- Simple object storage based on object type
- The right architecture means this can also serve as preservation layer w/ backups
PEOPLE to TRANSLATE FUNCTIONS into TECH

Project 1
- Text

Project 2
- Images
- Audio / Video

Project 3
- Data

Project 4
- Metadata

Administrative Layer
- Collect/Create/Organize
- “Data” deposit
- Metadata creation
- Customizable for each project

Life of a Project
**Simple Index /Access Layer**
- Basic search and browse
- Render in browser
- If all ‘project-specific’ access goes away, humans or machines still have granular access to content
Engagement / Innovation Layer
- Enables creative interaction with and re-use of data
- N-gram viewers
- Visualization tools
- Mapping tools
- Image recognition
- Project-specific website to point people to
Lifespans will vary for different layers

- The more innovative something is, the less likely it is to last a long time. This is an issue of balancing risk and reward.
- Many of the technologies under ‘basic access’ are governed by (open source) communities.
Owners, roles, and responsibilities can also vary across the service layers
Minimum Viable Service

- Project 1
- Project 2
- Project 3
- Project 4

Collect Create Organize

Storage and Preservation

- Text
- Images
- Audio / Video
- Data
- Metadata

Access / Discover

Lightweight software for object-level discovery and access

Basic Access

- Aggregated access
  - APIs / SOLR
  - Index
  - Individual Item Access
  - Image viewer
  - Video/audio player
  - Digital Edition Viewer

Engagement Analysis or Innovation

Present (Project websites)

- Image recognition
- Aggregated access
  - Text analysis (e.g. complex search, n-grams, etc)

PEOPLE to TRANSLATE FUNCTIONS into TECH

Life of a Project

- Project funding is gone
- PI has moved on
- ‘Data’ is still discoverable and useable at a granular level
THIS can still be rebuilt

Minimum Viable Service

Collect
Create
Organize

Project 2

Project 1

Project 3

Project 4

PEOPLE to TRANSLATE FUNCTIONS into TECH

Storage and Preservation

Access / Discover

Basic Access

Engagement
Analysis or
Innovation

Present
(Project websites)

Lightweight software for object-level discovery and access

Aggregated access

APIs / SOLR

Index

Text analysis (e.g. complex search, n-grams, etc)

Image recognition

Visualization

Compare

Present

Text

Images

Audio / Video

Data

Metadata

Image viewer

Video/ audio player

Digital Edition Viewer

Aggregated access

Individual Item Access

Basic Access

Life of a Project
iSicily corpus of Sicilian inscriptions

“...an open-ended, on-going, and highly collaborative project”
iSicily

- Collect
- Create
- Organize

Storage and Preservation

- Text
- Images
- Audio/Video
- Data
- Metadata

Access/Discover

- Light weighted software for object-level discovery and access

Basic Access

- Aggregated access
  - APIs
  - SOLR Index
- Individual Item Access
  - Image viewer
  - Video/audio player
  - Digital Edition Viewer

Engagement Analysis or Innovation

- Text analysis (e.g., complex search, n-grams, etc)
- Image recognition
- Mapping
- Compare
- Present

Present (Project websites)

iSicily Website
Minimum Viable Service

- Collect
- Create
- Organize

- Storage and Preservation
- Access / Discover
- Basic Access

- Aggregated access
  - APIs
  - SOLR Index
- Individual Item Access
  - Image viewer
  - Video/audio player
  - Digital Edition Viewer

- Basic Access

- Lightweight software for object-level discovery and access

- Engagement Analysis or Innovation

- Present (Project websites)

- Project funding is gone
- PI has moved on
- ‘Data’ is still discoverable and usable at a granular level

- Text analysis (e.g. complex search in grams, etc.)
- Image recognition
- Compare
- Present

- iSicily

- Website

- Text
- Images
- Audio / Video
- Data
- Metadata
Benefits

- Provides a minimum viable service
- Allows autonomy where it is needed
- Allows for different layers to have different lifespans and different owners
Risks

1. **The Funders**: Current funding models and funders specifically encourage technological innovation.

2. **The Perception**: Some projects may always insist that they cannot use a shared infrastructure due to their uniqueness.

3. **The Reality**: This modular, service-layer approach (or variations of it) may not easily accommodate the migration of *all* existing projects. With enough money all things are possible, but this may not be financially worthwhile.
Questions? Comments?

Do you have a similar approach? Let’s discuss.

madsen@athenaeum21.com
@mccarthymadsen

Athenaeum21 Consulting

www.athenaeum21.com
Project 2
Lightweight software for object-level discovery and access

Project 1
Text
Images
Audio / Video
Data
Metadata

Access / Discover

Project 3
Access

Basic Access

Project 4
Basic Access

Engagement Analysis or Innovation

Text analysis (e.g. complex search, n-grams, etc)

Visualization

Compare

Present

Present (Project websites)

Life of the funded project

Lifespan “forever”

Lifespan 5-10 years

Lifespan 3-5 years

Lifespan “forever”

Collect Create Organize

Storage and Preservation

APIs / SOLR
Index

Individual Item Access

Image viewer

Video / audio player

Digital Edition Viewer

PEOPLE to TRANSLATE FUNCTIONS into TECH

Collect Create Organize

Storage and Preservation

Access / Discover

Basic Access

Engagement Analysis or Innovation

Present (Project websites)

Collect Create Organize

Storage and Preservation

Access / Discover

Basic Access

Engagement Analysis or Innovation

Present (Project websites)

Collect Create Organize

Storage and Preservation

Access / Discover

Basic Access

Engagement Analysis or Innovation

Present (Project websites)
**Access / Discover**

- Aggregated access
  - APIs
  - SOLR Index

**Basic Access**

- Individual Item Access
  - Image Viewer
  - Video/audio player
  - Text viewer

**Engagement Analysis or Innovation**

- Text analysis (e.g. complex search, n-grams, etc)
  - Image recognition
  - Specialized Index
  - Compare
  - Present

**Collect Create Organize**

- Cult of Saints Website

**Storage and Preservation**

- Cult of Saints

**Access / Discover**

- Lightweight software for object-level discovery and access

**Basic Access**

- Text
  - Images
  - Audio/Video
  - Data

**Collect Create Organize**

- Metadata
Individual Item Access

Access / Discover

Aggregated access

Basic Access

Newton Project

Collect Create Organize

Storage and Preservation

Newton Project Website

Compare

Present (Project websites)

Engagement Analysis or Innovation

Text analysis (e.g. complex search, n-grams, etc)

APIs

SOLR Index

Newton Project Website

Image recognition

Specialized Index

Digital Edition Viewer

Video/audio player

Image viewer

Individual Item Access

Video/audio player

Digital Edition Viewer

Lightweight software for object-level discovery and access

Text

Images

Audio/Video

Data

Metadata

Collect Create Organize

Storage and Preservation

Access / Discover

Basic Access
Benefits

- Provides a minimum viable service
- Allows autonomy where it is needed
- Allows for different layers to have different lifespans and different owners
Risks

1. **The Funders**: Current funding models and funders specifically encourage technological innovation.

2. **The Perception**: Some projects may always insist that they cannot use a shared infrastructure due to their uniqueness.

3. **The Reality**: This modular, service-layer approach (or variations of it) may not easily accommodate the migration of *all* existing projects. With enough money all things are possible, but this may not be financially worthwhile.
Questions? Comments?

Do you have a similar approach? Let’s discuss.

madsen@athenaeum21.com
@mccarthymadsen
Athenaeum21 Consulting
www.athenaeum21.com