

Small, thick, and slow

Thinking about data and research publication in the
Humanities in the age of Open and FAIR

Daniel Paul O'Donnell
University of Lethbridge

Curtin University
November 25, 2019

About this paper

- Going to be speaking of how data are used in the humanities and implications for infrastructure design
 - How infrastructure currently interacts with typical humanities research practices
 - Why humanities researchers have been slow to adopt such infrastructure
 - How this infrastructure can be adapted to support (and improve) humanities research *without requiring it to abandon its primary features/strengths*
 - “Small” — focussed on very small number of data points or sets
 - “Thick” — involves intense curation and analysis of these few data
 - “Slow” — the same data points can be subject to years (generations) of subsequent, alternate, and supplementary analysis

About this paper

- Two parts:
 - 1: The problem of Humanities data
 - 2: A solution

About this paper

- Important to recognise that I'm dealing in generalities
 - Not all humanities data are small or “representational” in focus
 - Not all humanities work is about thick description
 - Not all humanities work is about reworking old material
- But much is and these are the ones that are least well catered to in current infrastructure

About me

- Traditionally trained medieval philologist and textual critic
- Means history of “big” and small data techniques
 - Thesis (1996) was analysis of (unpublished) database of textual variation in the Old English poetic canon
 - Letter-by-letter differences in about 20 poems surviving in more than one copy from the pre-conquest period
 - Later (2005) did 100,000 word edition of 9-line *Cædmon’s Hymn* (s. viii)
 - Now working on 5 object “edition” of the cross in pre-conquest England
- But
 - Coming from a textual/linguistic/literary approach
 - Focus on “editing” (i.e. the development and publication of “Primary Source” material — mediated representational data)

Part 1

The problem of humanities data

Traditionally, humanists resist speaking of data

- “Primary sources” = Texts, artifacts, objects of study
 - Can be originals (i.e. the artifact itself)
 - More often mediated and contextualised in some way (i.e. an edition, transcription, or similar)
- “Secondary sources” = Works of other scholars (often based on “Primary sources”)
- “Readings” (1) = Passages, extracts, quotations for interpretation or support
- “Readings” (2) = Interpretation, the end product of research (literary study)

Traditionally, humanists resist speaking of data

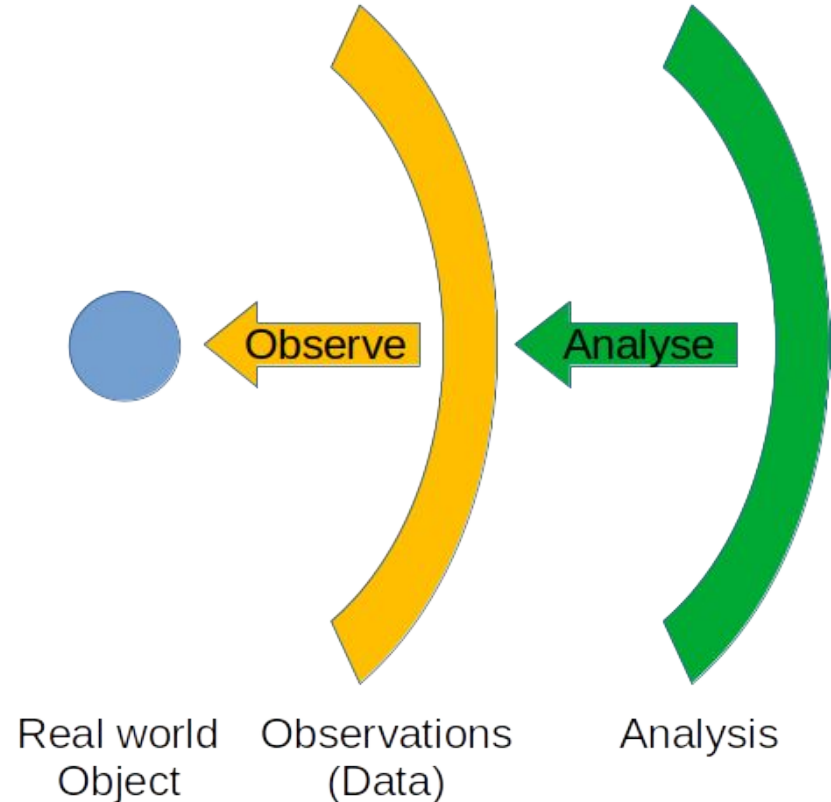
- These definitions are highly contingent
 - “Primary source” in one context can be the “secondary source” in another (and vice versa)
 - Or simultaneously “Primary” and “Secondary” (e.g. a critical edition)
- Also hard to constrain

“[a]lmost any document, physical artifact, or record or human activity can be used to study culture” and arguments proposing previously unrecognised sources (“high school yearbooks, cookbooks, or wear patterns in the floors of public places”) are valued acts of scholarship”

(Borgman 2007)

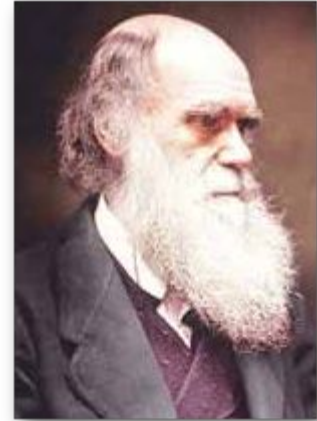
How does data work in other fields?

- Resistance makes sense, because Humanities data is different from other forms of data
- In other domains, “data” (“given things”) is more properly “capta” (“taken”): generated through experiment, observation, and measurement, then observed
- Think about Darwin and his work in the Galapagos Islands
 - What is his data?



How does data work in other fields?

- Resistance makes sense, because Humanities data is different from other forms of data
- In other domains, “data” (“given things”) is more properly “capta” (“taken”): generated through experiment, observation, and measurement, then observed
- Think about Darwin and his work in the Galapagos Islands
 - What is his data?



How does data work in other fields?

- Resistance makes sense, because Humanities data is different from other forms of data
- In other domains, “data” (“given things”) is more properly “capta” (“taken”): generated through experiment, observation, and measurement, then observed
- Think about Darwin and his work in the Galapagos Islands
 - What is his data?



The finches?

How does data work in other fields?

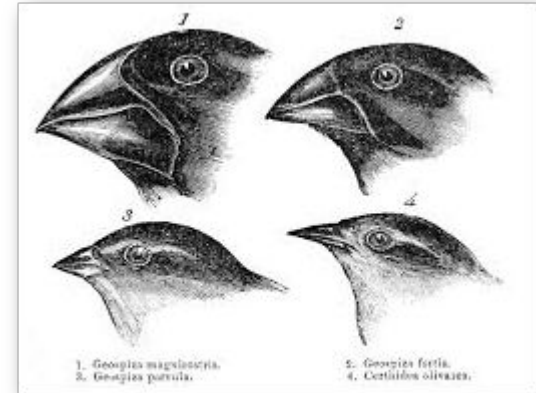
- Resistance makes sense, because Humanities data is different from other forms of data
- In other domains, “data” (“given things”) is more properly “capta” (“taken”): generated through experiment, observation, and measurement, then observed
- Think about Darwin and his work in the Galapagos Islands
 - What is his data?



The notes about the finches?

How does data work in other fields?

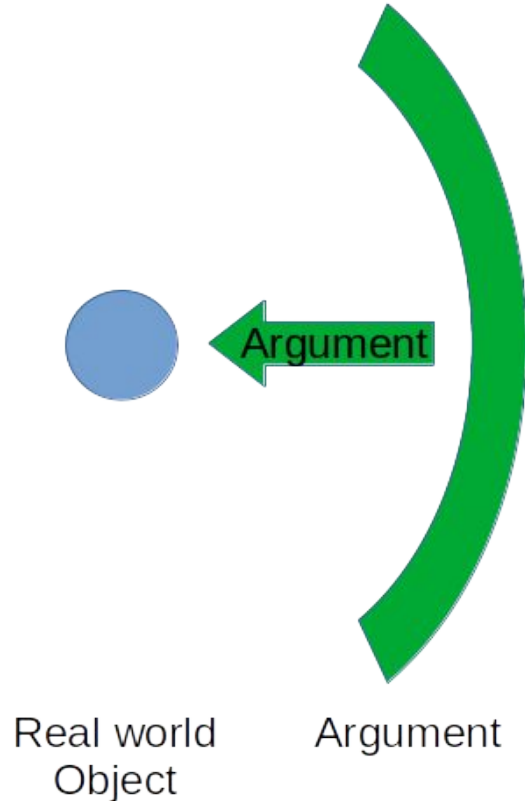
- In fact, in the sciences, it is the notes.
- “Data” = “represent[ation of] information in a formalized manner suitable for communication, interpretation, or processing” (NASA 2012); “the facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors” (NRC 1999)



The notes about the finches.

But in the humanities?

- Can be both “data” and “capta”, but very often “data”
- Very specific and often provisional: small;
- Depends on interpretation and argument (argue whether something *is* data): thick;
- Frequently revisit the same datasets to see them differently, provide new contexts, reuse: slow



But in the humanities?

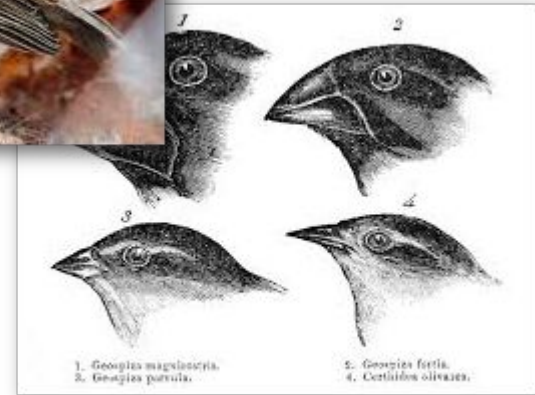
- Can be both “data” and “capta”, but very often “data”
- Very specific and often provisional: small;
- Depends on interpretation and argument (argue whether something *is* data): thick;
- Frequently revisit the same datasets to see them differently, provide new contexts, reuse: slow



Usually the Finch.

But in the humanities?

- Can be both “data” and “capta”, but very often “data”
- Very specific and often provisional: small;
- Depends on interpretation and argument (argue whether something *is* data): thick;
- Frequently revisit the same datasets to see them differently, provide new contexts, reuse: slow



Usually the Finch. Sometimes the notes.

But in the humanities?

- Can be both “data” and “capta”, but very often “data”
- Very specific and often provisional: small;
- Depends on interpretation and argument (argue whether something *is* data): thick;
- Frequently revisit the same datasets to see them differently, provide new contexts, reuse: slow



Usually the Finch. Sometimes the notes. And sometimes what Darwin thought he was doing in his notes about the Finch.

In Humanities, “Data” is arguably mostly “Finch”

- Interesting proof: Humanities “data,” unlike science “data” is almost all practically and theoretically non-rivalrous.
- Humanities researchers rarely have an incentive (or capability) to prevent others from accessing their raw material.
- 200 years of Jane Austen studies based on five main pieces of data.



Usually the Finch. Sometimes the notes. And sometimes what Darwin thought he was doing in his notes about the Finch.

The “Digital Humanities” don’t change this

- DH adds to this basic fact, but doesn’t change it:
 - We can now have “capta” (intermediate “observations” extracted algorithmically to form large data sets that then require interpretation)
 - We can now work across complete historical or geographic corpora: all known nineteenth-century English periodicals; every surviving tract from the U.S. Civil War
 - Introduces the possibility of deductive work
 - Makes method questions more important than when you worked inductively from the collections you could access

The “Digital Humanities” don’t change this

- But DH is not the perfection of the Humanities
 - A lot of research continues with “data” rather than “capta”
 - This “traditional” work remains sound and important
 - The distinction between “capta” and “data” is not teleological
 - “Big data” (“big capta”) DH is not better than “small data” (traditional) Humanities
 - Not all DH is “big capta” (you can do traditional work with computers)
 - “Big capta” approaches to Humanities questions can miss the point
 - Intensive curation and analysis of small data sets remains a major function of humanities research

Why does this matter?

- Although much humanities research is (appropriately) “small, thick, and slow,” it is also, in theory, useful for “big capta” work
 - Collectively, traditional humanists produce a lot of *very* high quality data
 - *Intensely* curated datasets and data points;
 - Broadly compatible with each other (i.e. each generation reedits and reconsiders the canon);
- If we could find a way to capture the value of this traditional data in a way that would allow them to be reused,
 - We’d have extremely useful material to repurpose
 - We’d be maximising the benefit of the traditional work that has been done on it

Why does this matter?

- But FAIR small data is by-and-large uneconomical for small data researchers
 - Their goal is to publish *contextualised* small-data datasets to
 - Serve as primary sources for others
 - e.g. an edition of Jane Austen's *Pride and Prejudice* is intended to support secondary work on that novel
 - Support very specific arguments about the specific instance
 - e.g. that there are three versions of *Hamlet*
 - The features that are required for reuse require (in essence) a separate, standalone, publication
 - Deposit in repository
 - Standardised metadata
 - Loss of key interpretative context and information

The case of manuscript photography

- Since the mid-1990s, there have been hundreds if not thousands of digital editions published of medieval and renaissance texts.
- Almost all of these contain high quality digital photographs of the original artifacts, often with very detailed, research-based expert commentary and analysis (transcriptions, bibliographic and other descriptions, etc.)
- Represents, *in theory*, a potentially huge, extremely rich, dataset for new cross-project work
 - Automatic scribe identification
 - Dating training sets
 - History of the Book

The case of manuscript photography

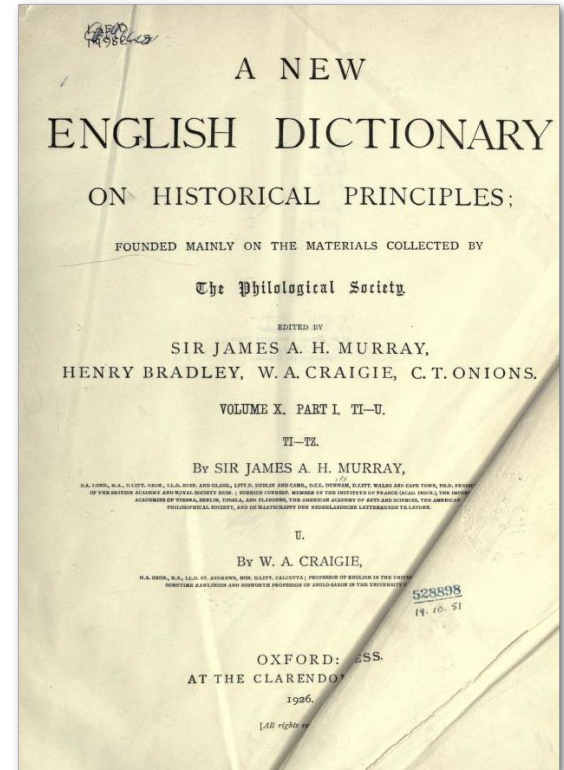
- Because the purpose of these photographs has been to support the contextual analysis and/or supply users with representations of the individual objects in question, very few are easily recovered or used by machines:
 - Few/no standards for metadata, APIs, etc
 - Very few explicitly connected to expert description
 - Relationship to other images and publication status not machine readable
- Result is a lost opportunity to create a “big capta” dataset of thickly described data from hundred of individual “small data” projects

So what to do?

- The solution to this is to accept the traditional nature and use-case involved in the production and consumption of Humanities research data
 - I.e. recognise that FAIR must accommodate the small, thick, and slow as easily as it does the big stand-alone examples from STEM
- That means that we have to either
 - Work within the traditional Humanities research workflow
 - Encourage traditional Humanities researchers to work within ours
- **As long as FAIR data publication means, in essence, publishing small, thick, and slow data twice (once in context and once without), we will never fully reap the benefit of these important and potentially huge cultural datasets**

We've been here before

- The New English Dictionary provides a non-digital model for this
 - Based on “historical principles” (i.e. definitions from and supported by historical quotations)
 - Massive crowd-sourced big-data collection, involving thousands collecting 1.8 million quotation slips from thousands of books prepared by generations of authors, scholars, and publishers (i.e. small data datasets)
 - In essence, an analogue version of what we want to do digitally

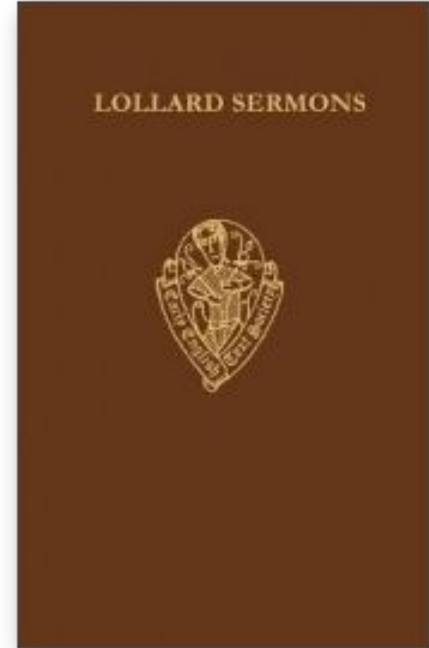


We've been here before

- They had the same problem
 - Discovered almost immediately after setting up the reading programme that the texts they were planning to use were unsuitable
 - Not available in modern editions
 - Poor or difficult-to-determine quality
 - In other words, they discovered that they needed to improve and standardise the small datasets from which they were going to draw their big data records.

We've been here before

- Solution? Create platform for new editions
 - Text societies and publishers to publish editions that met the NED's requirements
 - Encouraged leading scholars to edit (and later reedit) the texts they needed
 - Very symbiotic relationship between what was going on in historical textual research at the time and the needs of this big-data dictionary
- Result was an increase in high quality small-data editions *and* better big-data data set for NED



What to do

- What we need is something similar for the digital age
 - A workflow that encourages small-data researchers to prepare their datasets in a way that
 - Respects their traditional requirements for the intensive curation and analysis of individual data points or small datasets
 - Opens these small, thick, and slow datasets up to big data analysis
 - Does not increase (and preferably reduces) the cost of production, publication, and maintenance
- **A workflow in which suitability for “big capta” research is inherent in the publication “small data” workflow rather than a separate step.**

What to do?

- What we need, therefore, is a similar approach for the digital age, that is comfortable dealing with the small, thick, and slow nature of the work
 - Has to accept that most Humanities research is (properly) about a small numbers of objects (small)
 - That the purpose of most Humanities research is to analyse these small number of datapoints intensely (thick)
 - That researchers are going to want to rework these individual data points as part of the natural progress of their research (slow)
- **A workflow in which suitability for “big capta” research is inherent in the publication “small data” workflow rather than a separate step.**

Part 2

Being FAIR to the small, thick, and slow

Introduction

- In the rest of this talk, I'm going to talk about the “Data-First” approach we are developing for the Visionary Cross Project
 1. The project and some of our parameters
 2. Background issues and models
 3. The implementation
 4. Further work

About the Visionary Cross Project

- 9 year-old SSHRC funded project to produce an “edition” and “archive” of the “Visionary Cross cultural matrix” in Anglo-Saxon England
 - “Edition” means “Scholarly mediated reproduction”
 - “Archive” means “dataset of facsimiles and transcriptions”
 - “Visionary Cross Cultural Matrix” means “Collection of individual objects that also belong together for cultural reasons”

About the Visionary Cross Project

- Objects include some of the best known objects and texts from Pre-conquest England and Scotland.

About the Visionary Cross Project

- Objects include some of the best known objects and texts from Pre-conquest England and Scotland.



Vercelli Book Dream of the Rood and Elene poems (s. x/xi, South)

About the Visionary Cross Project

- Objects include some of the best known objects and texts from Pre-conquest England and Scotland.



Ruthwell Cross (s. VIII, North)

About the Visionary Cross Project

- Objects include some of the best known objects and texts from Pre-conquest England and Scotland.



Bewcastle Cross (s. viii, North)

About the Visionary Cross Project

- Objects include some of the best known objects and texts from Pre-conquest England and Scotland.



Brussels Cross (s. x/xi, South)

About the Visionary Cross Project

- Interesting as individual objects and as a group:
 - Span period temporally, geographically, linguistically
 - (possibly) Earliest attested poetry
 - Complete runic poem
 - Include 1 of only 2~3 examples of poetic quotation
 - “Multiply attested” poetic text (>3% of the corpus)
 - Related to each other thematically (cult of the cross) and textually and/or artistically

About the Visionary Cross Project

- In other words we anticipate use as both
 - A traditional small-data project (as well as a not-so-traditional small-data project):
 - Individuals coming to us for limited amounts of data in the context of our thick description because they want to use our material as the primary source for subsequent work
 - A contribution to potential big-data purposes:
 - Data that can be used, reused, supplemented, and aggregated by others without negotiation

Project Requirements

- A. Flexible:
 - Choose to view individual/group in appropriate format
- B. Extensible:
 - Add, rearrange, or reuse material without negotiation
- C. Authoritative:
 - Preserve credit/responsibility for all contributions
- D. Durable:
 - Permanently discoverable and available
 - Low/no maintenance

Different approaches over the years

- Wiki?
 - Flexible (e.g. categories/entries) (**A**)
 - Add and (re)connect material without negotiation (**B**)
 - But
 - Doesn't preserve Authority (**G**)
 - Requires ongoing maintenance (**D**)
 - One kind of presentation (**A**)

Different approaches over the years

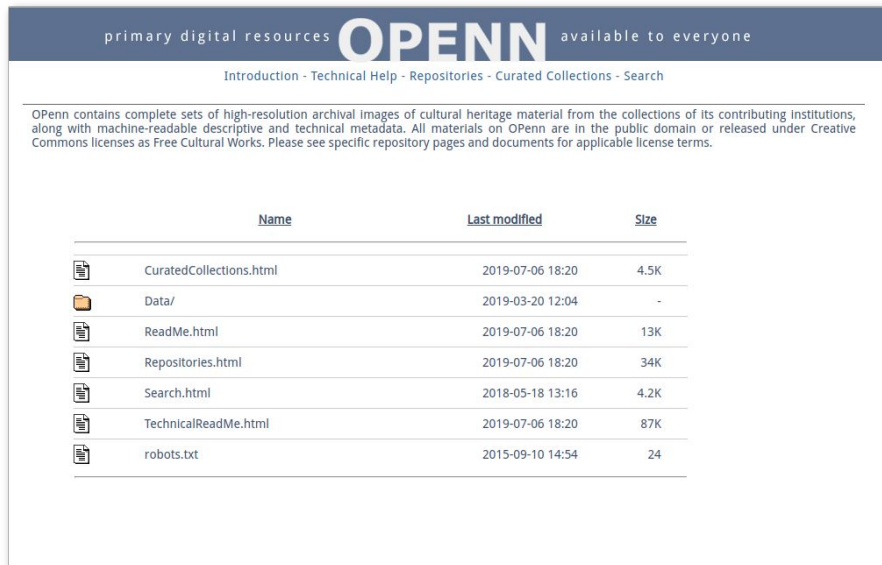
- Game engine
 - Provided different ways of organising material and good at object/collection (**A**)
 - Preserved authority (**C**)
 - Some engines allowed some external contributions (**B**)
 - But
 - Requires others to use our system (**B**)
 - None strong on external contributions (**B**)
 - Requires ongoing maintenance (**D**)

OPenn (<http://openn.library.upenn.edu/>)








- Repository for MS information, images, transcriptions
- Built to replace a previous “turning the pages” type interface for MS collections
 - Open the collection up to machine access (i.e. via rsync, ssh, ftp, etc)
 - Maintain human readability

OPenn (<http://openn.library.upenn.edu/>)

- Essentially a lightly-skinned directory structure (i.e. a RESTful-like API)
 - Human-readable HTML pages

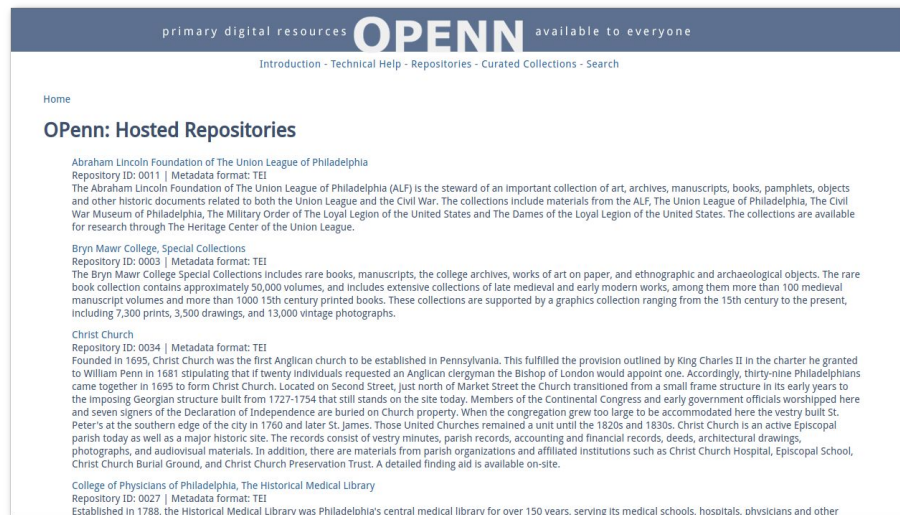


The screenshot shows the OPenn website interface. At the top, there is a dark blue header with the text "primary digital resources" on the left, the "OPENN" logo in the center, and "available to everyone" on the right. Below the header is a navigation bar with links: "Introduction - Technical Help - Repositories - Curated Collections - Search". A paragraph of text states: "OPenn contains complete sets of high-resolution archival images of cultural heritage material from the collections of its contributing institutions, along with machine-readable descriptive and technical metadata. All materials on OPenn are in the public domain or released under Creative Commons Licenses as Free Cultural Works. Please see specific repository pages and documents for applicable license terms." Below this text is a table with three columns: "Name", "Last modified", and "Size". The table lists several files and folders with their respective last modified dates and sizes.

	Name	Last modified	Size
	CuratedCollections.html	2019-07-06 18:20	4.5K
	Data/	2019-03-20 12:04	-
	ReadMe.html	2019-07-06 18:20	13K
	Repositories.html	2019-07-06 18:20	34K
	Search.html	2018-05-18 13:16	4.2K
	TechnicalReadMe.html	2019-07-06 18:20	87K
	robots.txt	2015-09-10 14:54	24

OPenn (<http://openn.library.upenn.edu/>)

- Essentially a lightly-skinned directory structure (i.e. a RESTful-like API)
 - Human-readable HTML pages



The screenshot shows the OPenn website interface. At the top, there is a navigation bar with the text "primary digital resources" on the left, the "OPENN" logo in the center, and "available to everyone" on the right. Below the navigation bar, there is a menu with "Introduction - Technical Help - Repositories - Curated Collections - Search". The main content area is titled "Home" and "OPenn: Hosted Repositories". It lists three repositories: 1. Abraham Lincoln Foundation of The Union League of Philadelphia (Repository ID: 0011), 2. Bryn Mawr College Special Collections (Repository ID: 0003), and 3. Christ Church (Repository ID: 0034). Each repository entry includes a brief description of the collection and its historical significance.

primary digital resources **OPENN** available to everyone

Introduction - Technical Help - Repositories - Curated Collections - Search

Home

OPenn: Hosted Repositories

Abraham Lincoln Foundation of The Union League of Philadelphia
Repository ID: 0011 | Metadata format: TEI
The Abraham Lincoln Foundation of The Union League of Philadelphia (ALF) is the steward of an important collection of art, archives, manuscripts, books, pamphlets, objects and other historic documents related to both the Union League and the Civil War. The collections include materials from the ALF, The Union League of Philadelphia, The Civil War Museum of Philadelphia, The Military Order of The Loyal Legion of the United States and The Dames of the Loyal Legion of the United States. The collections are available for research through The Heritage Center of the Union League.

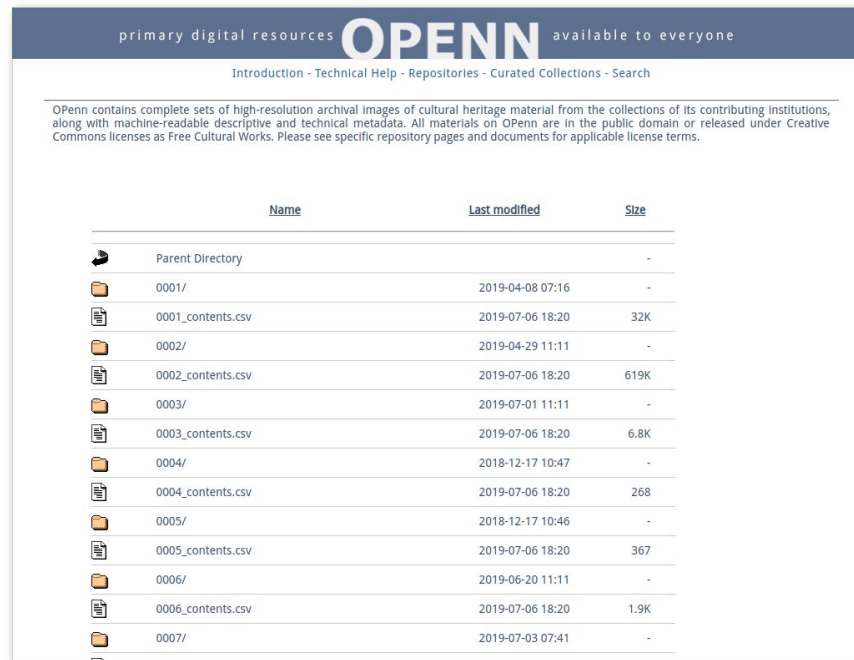
Bryn Mawr College Special Collections
Repository ID: 0003 | Metadata format: TEI
The Bryn Mawr College Special Collections includes rare books, manuscripts, the college archives, works of art on paper, and ethnographic and archaeological objects. The rare book collection contains approximately 50,000 volumes, and includes extensive collections of late medieval and early modern works, among them more than 100 medieval manuscript volumes and more than 1000 15th century printed books. These collections are supported by a graphics collection ranging from the 15th century to the present, including 7,300 prints, 3,500 drawings, and 13,000 vintage photographs.

Christ Church
Repository ID: 0034 | Metadata format: TEI
Founded in 1695, Christ Church was the first Anglican church to be established in Pennsylvania. This fulfilled the provision outlined by King Charles II in the charter he granted to William Penn in 1681 stipulating that if twenty individuals requested an Anglican clergyman the Bishop of London would appoint one. Accordingly, thirty-nine Philadelphians came together in 1695 to form Christ Church. Located on Second Street, just north of Market Street the Church transitioned from a small frame structure in its early years to the imposing Georgian structure built from 1727-1754 that still stands on the site today. Members of the Continental Congress and early government officials worshipped here and seven signers of the Declaration of Independence are buried on Church property. When the congregation grew too large to be accommodated here the vestry built St. Peter's at the southern edge of the city in 1760 and later St. James. Those United Churches remained a unit until the 1820s and 1830s. Christ Church is an active Episcopal parish today as well as a major historic site. The records consist of vestry minutes, parish records, accounting and financial records, deeds, architectural drawings, photographs, and audiovisual materials. In addition, there are materials from parish organizations and affiliated institutions such as Christ Church Hospital, Episcopal School, Christ Church Burial Ground, and Christ Church Preservation Trust. A detailed finding aid is available on-site.















College of Physicians of Philadelphia, The Historical Medical Library
Repository ID: 0027 | Metadata format: TEI
Established in 1788, the Historical Medical Library was Philadelphia's central medical library for over 150 years, serving its medical schools, hospitals, physicians and other

OPenn (<http://openn.library.upenn.edu/>)

- Essentially a lightly-skinned directory structure (i.e. a RESTful-like API)
 - Human-readable HTML pages



The screenshot shows the OPenn website interface. At the top, it says "primary digital resources OPENN available to everyone". Below that, there are navigation links: "Introduction - Technical Help - Repositories - Curated Collections - Search". A paragraph of text explains that OPenn contains complete sets of high-resolution archival images of cultural heritage material from the collections of its contributing institutions, along with machine-readable descriptive and technical metadata. All materials on OPenn are in the public domain or released under Creative Commons licenses as Free Cultural Works. Below this text is a table listing directory entries.

	<u>Name</u>	<u>Last modified</u>	<u>Size</u>
	Parent Directory		-
	0001/	2019-04-08 07:16	-
	0001_contents.csv	2019-07-06 18:20	32K
	0002/	2019-04-29 11:11	-
	0002_contents.csv	2019-07-06 18:20	619K
	0003/	2019-07-01 11:11	-
	0003_contents.csv	2019-07-06 18:20	6.8K
	0004/	2018-12-17 10:47	-
	0004_contents.csv	2019-07-06 18:20	268
	0005/	2018-12-17 10:46	-
	0005_contents.csv	2019-07-06 18:20	367
	0006/	2019-06-20 11:11	-
	0006_contents.csv	2019-07-06 18:20	1.9K
	0007/	2019-07-03 07:41	-

OPenn (<http://openn.library.upenn.edu/>)

- Essentially a lightly-skinned directory structure (i.e. a RESTful-like API)
 - Human-readable HTML pages

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<TEI xmlns="http://www.tei-c.org/ns/1.0">
  ▼<teiHeader>
    ▼<fileDesc>
      ▼<titleStmt>
        ▼<title>
          Description of University of Pennsylvania LJS
          16: Speculum historiale, Books 25-28
        </title>
      </titleStmt>
      ▼<publicationStmt>
        <publisher>The University of Pennsylvania
        Libraries</publisher>
      ▼<availability>
        ▼<licence
          target="http://creativecommons.org/licenses/by/4.0/"
          This description is ©2015 University of
          Pennsylvania Libraries. It is licensed under
          a Creative Commons Attribution License
          version 4.0 (CC-BY-4.0)
          https://creativecommons.org/licenses/by/4.0/
          For a description of the terms of use see
          the Creative Commons Deed
          https://creativecommons.org/licenses/by/4.0/
          /license
```


OPenn (<http://openn.library.upenn.edu/>)

- Essentially a lightly-skinned directory structure (i.e. a RESTful-like API)
 - Human-readable HTML pages

```
document_id,path,title,metadata_type,created,updated
1,0001/ljs103,Reproduction of Sieneese book covers.,TEI,2014-11-03T23:38:42+00:00,2015-04-22T15:17:05+00:00
2,0001/ljs201,Evangelista Torricelli letter to Marin Marse
2014-11-03T23:39:46+00:00,2015-04-22T15:17:06+00:00
3,0001/ljs255,Manuscript leaf from De casibus virorum illu
2014-11-03T23:39:46+00:00,2015-04-22T15:17:06+00:00
4,0001/ljs489,Nawaz letter with seal,TEI,2014-11-03T23:40:
5,0001/ljsmisc1,Sluby family indenture :,TEI,2014-11-03T23
6,0001/ljsmisc2,Timothy Stedham indenture :,TEI,2014-11-03
7,0001/ljsmisc3,John and Mary Hoffman indenture :,TEI,2014
8,0001/ljsmisc4,Jacob Richman survey :,TEI,2014-11-03T23:4
9,0001/ljsmisc5,Subscription for cutting a channel from Sa
2014-11-03T23:43:59+00:00,2015-04-22T15:17:07+00:00
10,0001/ljsmisc6,John Hunt indenture :,TEI,2014-11-03T23:4
13,0001/ljs266,La generacion de Adam,TEI,2014-11-04T18:58:
15,0001/ljs204,Shesh kenafayim,TEI,2014-11-04T20:52:22+00:
16,0001/ljs491,Perush sefer ha-yesodot shel Uklids,TEI,201
17,0001/ljs217,Rapport sur les poids et mesures,TEI,2014-1
18,0001/ljs454,Seiyo senpaku zukai,TEI,2014-11-04T21:20:21
19,0001/ljs488,Cose di geometria,TEI,2014-11-04T21:32:36+0
20,0001/ljs496,Treatise on practical mathematical calculat
2014-11-04T21:35:40+00:00,2015-04-22T15:17:07+00:00
```

OPenn (<http://openn.library.upenn.edu/>)

- Love approach because it touches on all parts of vision
 - Flexible (i.e. **A**): can skin different groupings, focus on individuals or collections
 - Extensible (i.e. **B**): can extract from system
 - Authoritative (i.e. **C**): preserves authority
 - Durable (i.e. **D**): requires no software maintenance

OPenn (<http://openn.library.upenn.edu/>)

- But not perfect
 - Inflexible (i.e. **A**): Hierarchical data structure (can't have machine readable virtual collections)
 - Not extensible (i.e. **B**):
 - Additions/reorganisations require server access
 - Collections are “official” (entire libraries/fonds)
 - Not durable (i.e. **D**):
 - Publisher responsible for maintaining server
 - No persistent identifiers

Requirements (further points)

- E. Externally registered persistent identifiers
- F. Users need to be able to present alternatives/additions to our material inside or outside the same system
- G. Has to be “Publish-and-Forget”: once we are finished with it, it needs to be maintained by others.

Our solution

- Use Zenodo and GitHub to create an OPenn-like data repository, while answering its lacunae
- A “Data-first” approach to publication that
 1. Is human and machine readable
 2. Preserves attribution
 3. Open to non-negotiated addition, reorganisation, reuse
 4. Uses standard, third-party-maintained, persistent IDs
 5. Maintained for free by others (requires no post-publication maintenance by the project)

Zenodo

- EU-funded OpenAire Data Repository
 - Hosted at CERN
 - Guaranteed by EU
 - Accepts “all research outputs from all fields of science”
 - Assigns DOIs to all submissions (“conceptual” and “record”)
 - Based on Invenio Digital Repository Engine
 - Excellent metadata and LOD capabilities

GitHub

- Code repository, version control, distribution system
- Used by millions for developing code-based projects
- Recently added ability to publish web-pages using Jekyll-based “GitHub pages”
- Based on Open Source Git
- But
 - Recently bought by Microsoft (it’s always been private)
 - Not archival (conditions of use allow for suspension of service for any reason at any time)

Interaction of Zenodo and GitHub

- **GitHub repositories can be archived in Zenodo**
 - Snapshots are deposited in Zenodo as Zipped directories
 - Given a Zenodo DOI and treated like any other record
- Means:
 1. Replace GitHub's non-guarantee with Zenodo's permanent guarantee
 2. Presentation (versions) are also citable research objects (FAIR data AND FAIR code)

An example:

Cædmon's Hymn

- Originally CD-ROM (2005)
- Now online (2018)
- Code published using GitHub pages
 - <https://caedmon.seenet.org/>
 - <https://seenet-medieval.github.io/caedmonshymn>
- Code base preserved as Zenodo object (in all versions)

Cædmon's Hymn
A multimedia study, edition and archive

● Daniel Paul O'Donnell

With the assistance of Dawn Collins, Matt Van Egmond, Jon Lane, Catherine Larson, Angela Mlynarski, Asia Nelson, and Lyndon Simmons (2005) and Titilola Babalola Aiyegbusi and Gurpreet Singh (2018).

Version 1.1
Internet Reprint

TEI P4 SGML, and XHTML, 1.0 Transitional Conformant Edition

SEENET, A.8

Cambridge: D.S. Brewer, 2005
SEENET, 2018

DOI (code): [10.5281/zenodo.1198856](https://doi.org/10.5281/zenodo.1198856)
[How to cite this edition](#)

An example: *Cædmon's Hymn*

- Originally CD-ROM (2005)
- Now online (2018)
- Code published using GitHub pages
 - <https://caedmon.seenet.org/>
 - <https://seenet-medieval.github.io/caedmonshymn>
- Code base preserved as Zenodo object (in all versions)

The screenshot shows a Zenodo repository page for the project "Cædmon's Hymn: A multimedia study, edition, and archive internet edition. Version 1.1". The page is dated April 21, 2018, and has 86 views and 7 downloads. The author is Daniel Paul O'Donnell. The project is available for reading online at <http://caedmon.seenet.org/>. The project description mentions an online republication of Daniel Paul O'Donnell's "Cædmon's Hymn: A multimedia study, archive and edition" (Cambridge: D. S. Brewer, SEENET, and The Medieval Academy, 2005). The official release of version 1.1 is noted. A file browser shows a directory structure for "caedmonshymn-v1.1.zip" containing files like "gitignore", "travis.yml", "CNAME", "Gemfile", "LICENSE", "README.md", "fonts" (with sub-files like "Junicode-Bold.ttf", "Junicode-Bold.woff", etc.), and "htm". The total size of the files is 133.0 MB. The page also features an OpenAIRE logo, a publication date of April 21, 2018, a DOI of 10.5281/zenodo.1226549, and related identifiers including the project's GitHub repository and Zenodo tree.

An example: *Cædmon's Hymn*

- Originally CD-ROM (2005)
- Now online (2018)
- Code published using GitHub pages
 - <https://caedmon.seenet.org/>
 - <https://seenet-medieval.github.io/caedmonshymn>
- Code base preserved as Zenodo object (in all versions)

The image shows a Zenodo record for 'Cædmon's Hymn'. The record is displayed in a light grey box with a white background. The metadata includes the publication date (April 21, 2018), DOI (10.5281/zenodo.1226549), and keywords (Caedmon, Old English, scholarly edition). The imprint is the Society for Early English and Norse Electronic Texts, Raleigh, NC. The related identifiers include the URL http://caedmon.seenet.org and the GitHub repository https://github.com/seenet-medieval/caedmonshymn/tree/v1.1. The communities listed are Daniel Paul O'Donnell Personal Repository. The license is 'Other (Open)'. To the right of the record, there are statistics: 86 views and 7 downloads, a circular progress indicator showing 1 tweet, and the OpenAIRE logo. Below the record, there is a file download button labeled 'Files (133.0 MB)'. The background of the screenshot shows a blue header with the user email 'singhg@uleth.ca'.

Publication date:
April 21, 2018

DOI:
DOI 10.5281/zenodo.1226549

Keyword(s):
Caedmon Old English scholarly edition

Imprint:
Society for Early English and Norse Electronic Texts, Raleigh, NC.

Related identifiers:
Compiles:
<http://caedmon.seenet.org>
Identical to:
<https://github.com/seenet-medieval/caedmonshymn/tree/v1.1>

Communities:
Daniel Paul O'Donnell Personal Repository

License (for files):
[Other \(Open\)](#)

86 views 7 downloads
See more details...

1 Tweeted by 1
See more details

Indexed in
OpenAIRE

Publication date:
April 21, 2018

DOI:
DOI 10.5281/zenodo.1226549

Keyword(s):
Caedmon Old English scholarly edition

Imprint:
Society for Early English and Norse Electronic Texts, Raleigh, NC.

Related identifiers:
Compiles:
<http://caedmon.seenet.org>
Identical to:
<https://github.com/seenet-medieval/caedmonshymn/tree/v1.1>

Communities:
Daniel Paul O'Donnell Personal Repository

Files (133.0 MB)

An example: *Cædmon's Hymn*

- Originally CD-ROM (2005)
- Now online (2018)
- Code published using GitHub pages
 - <https://caedmon.seenet.org/>
 - <https://seenet-medieval.github.io/caedmonshymn>
- Code base preserved as Zenodo object (in all versions)

Publication date:
April 21, 2018

DOI:
DOI [10.5281/zenodo.1226549](https://doi.org/10.5281/zenodo.1226549)

Keyword(s):
[Caedmon](#) [Old English](#) [scholarly edition](#)

Imprint:
Society for Early English and Norse E
Texts, Raleigh, NC.

Related identifiers:
Compiles:
<http://caedmon.seenet.org>
Identical to:
<https://github.com/seenet-medieval/caedmonshymn/tree/v1.1>

Communities:
[Daniel Paul O'Donnell Personal Repos](#)

License (for files):
[Other \(Open\)](#)

Files (1330 MB)

Versions

Version v1.1	Apr 21, 2018
10.5281/zenodo.1226549	
Version v1.1-beta.05	Apr 18, 2018
10.5281/zenodo.1220238	
Version v1.1-beta.04	Mar 31, 2018
10.5281/zenodo.1210588	
Version v1.1-beta.03	Mar 30, 2018
10.5281/zenodo.1210005	
Version v1.1-beta.02	Mar 14, 2018
10.5281/zenodo.1198862	

[View all 6 versions](#)

Cite all versions? You can cite all versions by using the DOI [10.5281/zenodo.1198856](https://doi.org/10.5281/zenodo.1198856). This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)

Visionary Cross as Data

- Combining two systems allows us to publish a data-centric edition that is
 - Flexible
 - Extensible
 - Authoritative
 - Durable
 - Externally registered persistent IDs
 - Maintained by others

Heart is the Zenodo record

- Basic unit of edition (1 record = 1 datum)
- Provides machine readability, extensibility, persistence, and archiving
- *Also acts as document server for rest of edition

Zenodo record

- Human and machine readable metadata record + file(s)
- *Typed “additional identifiers”
- *Two kinds of DOIs:
 - “Conceptual” (latest)
 - “Version” (current)
- *RESTful files URLs
 - No link rot

The screenshot shows a Zenodo record page for 'Ruthwell Cross - 3D Model Ambient (15M)'. The page is dated April 24, 2018, and is associated with the 'The Visionary Cross' community. The record includes a DOI of 10.5281/zenodo.1490879 and a list of files: an XML record, a 3D model, and a 2D thumbnail. The record has 77 views and 15 downloads. It is indexed in OpenAIRE. The publication date is April 24, 2018. The keyword is 'Ruthwell Cross, The Visionary Cross Project, 3D Cultural Heritage, Digital Cultural Heritage, Anglo-Saxon, 3D Model, 3DHOP'. The record is documented by 10.5281/zenodo.1490863 and is licensed under Creative Commons Attribution 4.0 International.

zenodo Search Upload Communities singhg@uleth.ca

April 24, 2018 Dataset Open Access Edit

Ruthwell Cross - 3D Model Ambient (15M)

Singh, Gurpreet; O'Donnell, Daniel; The Visionary Cross Project

This is a High-Resolution Ambient 3D model the Ruthwell Cross, developed as part of the ongoing Visionary Cross project. This model has a 15M poly count.

This record contains:

- An xml record: `_Ruthwell_3DModel_Ambient_15M_Metadata.xml`
- A 3D Model of the Ruthwell cross: `cross_AmbientOcclusion_15M.ply`
- A 2D thumbnail: `Ruthwell_Cross00.png`

The DOI for this version of the record is 10.5281/zenodo.1490879. The DOI 10.5281/zenodo.1490878 always points to the latest version of this record.

This file in .ply format is best viewed using 3DHOP but can also be viewed using 3D rendering software like MeshLab.

For individual panels and other related material go to The Visionary Cross community at: https://zenodo.org/communities/the_visionary_cross/

Preview

Communities: The Visionary Cross Remove

77 views 15 downloads See more details...

Indexed in OpenAIRE

Publication date: April 24, 2018

DOI: 10.5281/zenodo.1490879

Keyword(s): Ruthwell Cross, The Visionary Cross Project, 3D Cultural Heritage, Digital Cultural Heritage, Anglo-Saxon, 3D Model, 3DHOP

Related identifiers: Documented by: 10.5281/zenodo.1490863

Communities: The Visionary Cross

License (for files): Creative Commons Attribution 4.0 International

Zenodo record

- Human and machine readable metadata record + file(s)
- *Typed “additional identifiers”
- *Two kinds of DOIs:
 - “Conceptual” (latest)
 - “Version” (current)
- *RESTful files URLs
 - No link rot

The screenshot shows the Zenodo record creation interface. At the top, there are three radio button options for access: Embargoed Access, Restricted Access, and Closed Access. Below these is a 'License' field with a dropdown menu set to 'Creative Commons Attribution 4.0 International'. A note below the license field states: 'Required. Selected license applies to all of your files displayed on the top of the form. If you want to upload some of your files under different licenses, please do so in separate uploads. If you cannot find the license you're looking for, include a relevant LICENSE file in your record and choose one of the Other licenses available (Other (Open), Other (Attribution), etc.). The supported licenses in the list are harvested from opendefinition.org and spdx.org. If you think that a license is missing from the list, please contact us.'

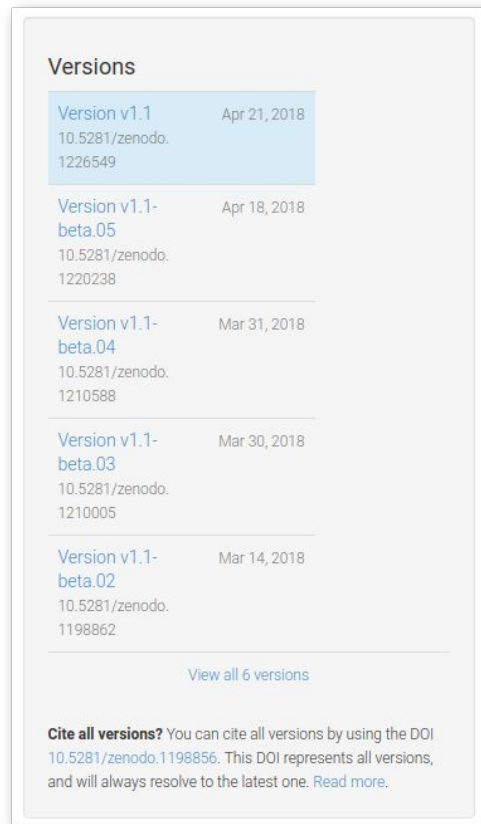
The 'Funding' section is titled 'recommended' and contains a text input for the funding agency name and a dropdown for 'Grants' with 'European Commission (EU)' selected. A note below states: 'Optional. OpenAIRE-supported projects only. For other funding acknowledgements, please use the Additional Notes field. Note: a human Zenodo curator will need to validate your upload - you may experience a delay before it is available in OpenAIRE.' There is a '+ Add another grant' button.

The 'Related/alternate identifiers' section is also titled 'recommended' and contains a text input for a related identifier (10.5281/zenodo.1490863) and a dropdown menu with 'documents this upload' selected. A note above the input states: 'Specify identifiers of related publications and datasets. Supported identifiers include: DOI, Handle, ARK, PURL, ISSN, ISBN, PubMed ID, PubMed Central ID, ADS Bibliographic Code, arXiv, Life Science Identifiers (LSID), EAN-13, ISTC, URNs and URLs.' There is a '+ Add another related identifier' button.

At the bottom, there is a list of optional fields: Contributors, References, Journal, Conference, Book/Report/Chapter, Thesis, and Subjects, each with a right-pointing arrow.

Zenodo record

- Human and machine readable metadata record + file(s)
- *Typed “additional identifiers”
- *Two kinds of DOIs:
 - “Conceptual” (latest)
 - “Version” (current)
- *RESTful files URLs
 - No link rot



The screenshot displays a Zenodo record page titled "Versions". It lists five versions of a dataset, each with a version number, a date, and a DOI. The first version, "Version v1.1", is highlighted in blue. Below the list is a link to "View all 6 versions". At the bottom, there is a section titled "Cite all versions?" which explains that a specific DOI (10.5281/zenodo.1198856) can be used to cite all versions, and it will always resolve to the latest one. A "Read more" link is provided for further information.

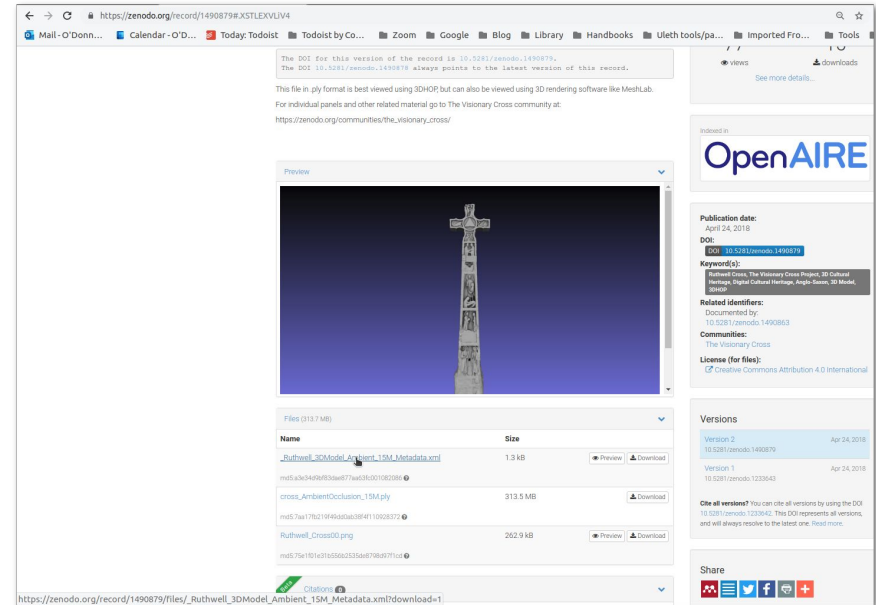
Versions	
Version v1.1 10.5281/zenodo. 1226549	Apr 21, 2018
Version v1.1- beta.05 10.5281/zenodo. 1220238	Apr 18, 2018
Version v1.1- beta.04 10.5281/zenodo. 1210588	Mar 31, 2018
Version v1.1- beta.03 10.5281/zenodo. 1210005	Mar 30, 2018
Version v1.1- beta.02 10.5281/zenodo. 1198862	Mar 14, 2018

[View all 6 versions](#)

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.1198856. This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)

Zenodo record

- Human and machine readable metadata record + file(s)
- *Typed “additional identifiers”
- *Two kinds of DOIs:
 - “Conceptual” (latest)
 - “Version” (current)
- *RESTful files URLs
 - No link rot

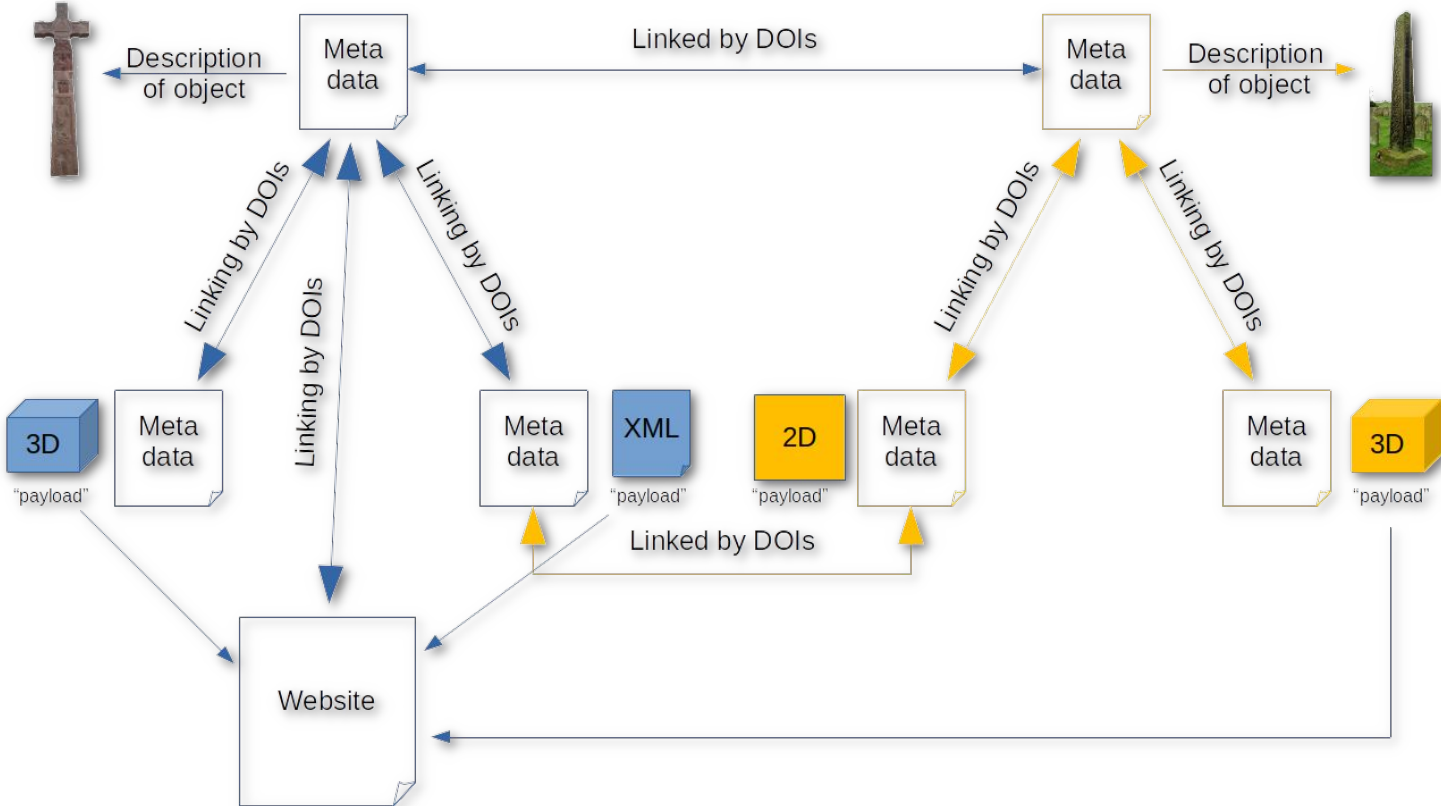


The screenshot shows a Zenodo record page for a 3D model of a cross. The URL is <https://zenodo.org/record/1490879#XSTLEXVLIV4>. The page displays a preview of the 3D model, which is a white cross against a blue background. Below the preview is a table of files:

Name	Size	Actions
Ruthwell_3DModel_Ambient_15M_Metadata.xml	1.3 kB	Preview Download
md5a3c34d49d83d4e77aa33c0010e2086		
cross_AmbientOcclusion_15M.ply	313.5 MB	Download
md57aa178d19f49a8a3a284f11028377		
Ruthwell_Cross00.png	262.9 kB	Preview Download
md575a1f01e31e556a2e35a678b8971cc		

The right sidebar contains metadata including the publication date (April 24, 2018), DOI (10.5281/zenodo.1490879), keywords (Ruthwell Cross, The Visionary Cross Project, 3D Cultural Heritage, Digital Cultural Heritage, Anglo-Saxon, 3D Model, JSON), related identifiers, and license (Creative Commons Attribution 4.0 International). The bottom of the page shows a share section with social media icons and a citation button.


Edition is built around records



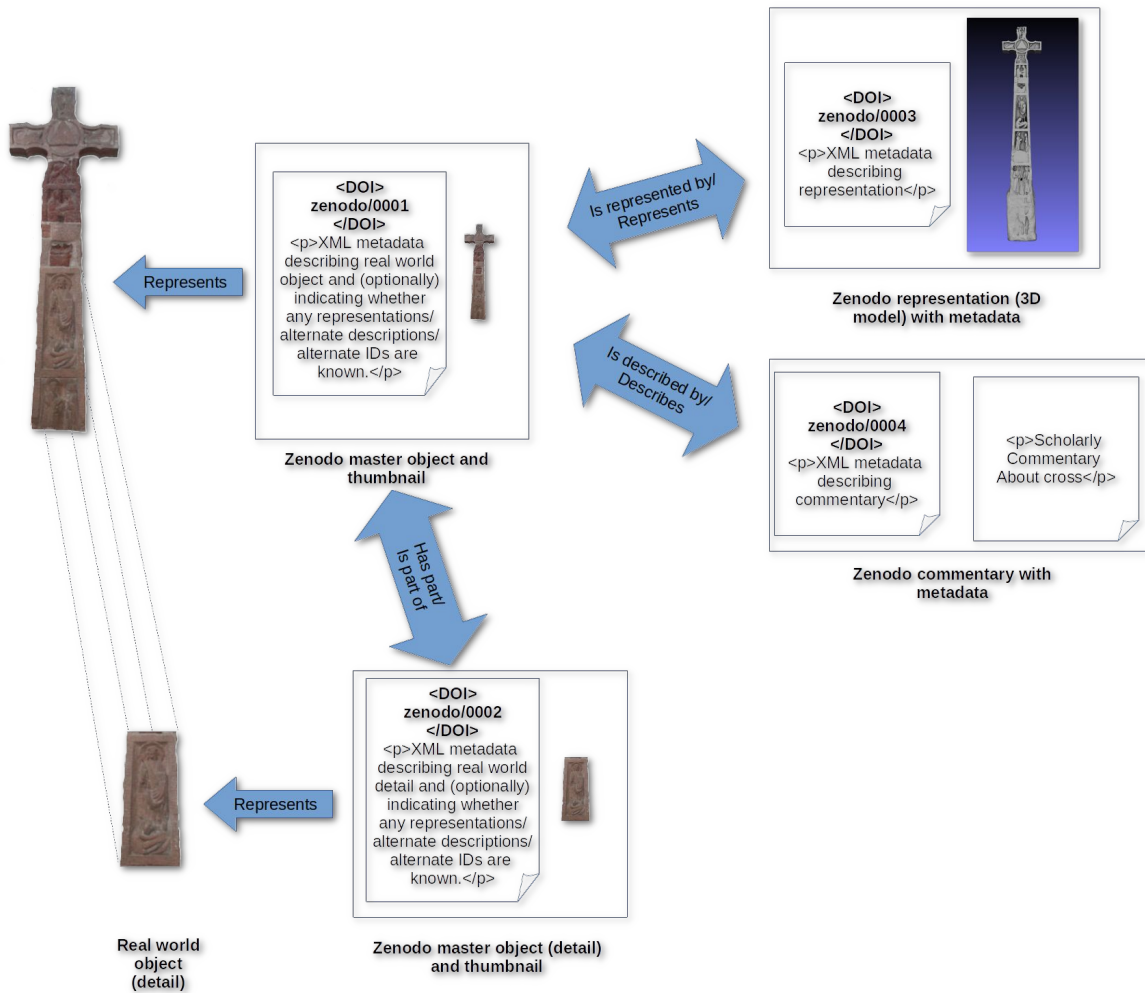


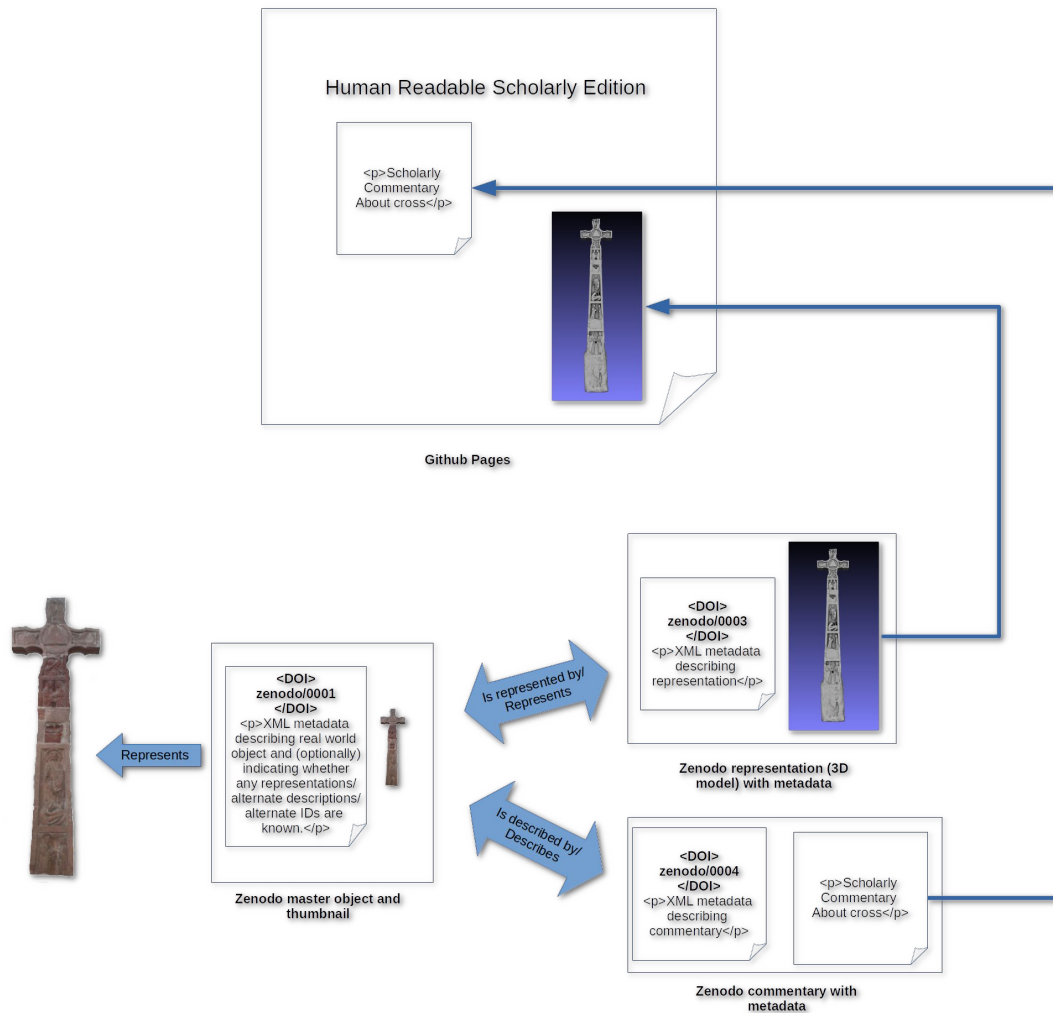
<DOI>
zenodo/0001
</DOI>

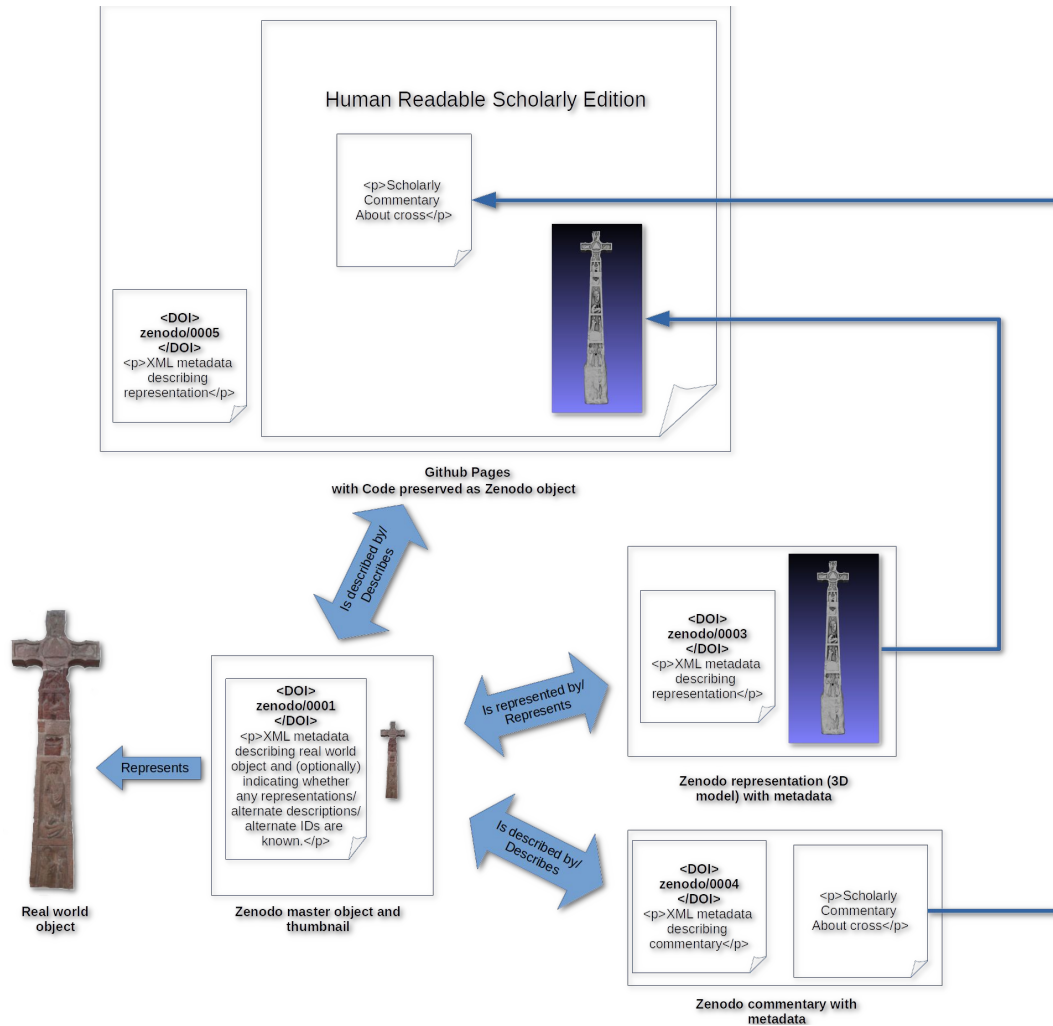
<p>XML metadata describing real world object and (optionally) indicating whether any representations/alternate descriptions/alternate IDs are known.</p>

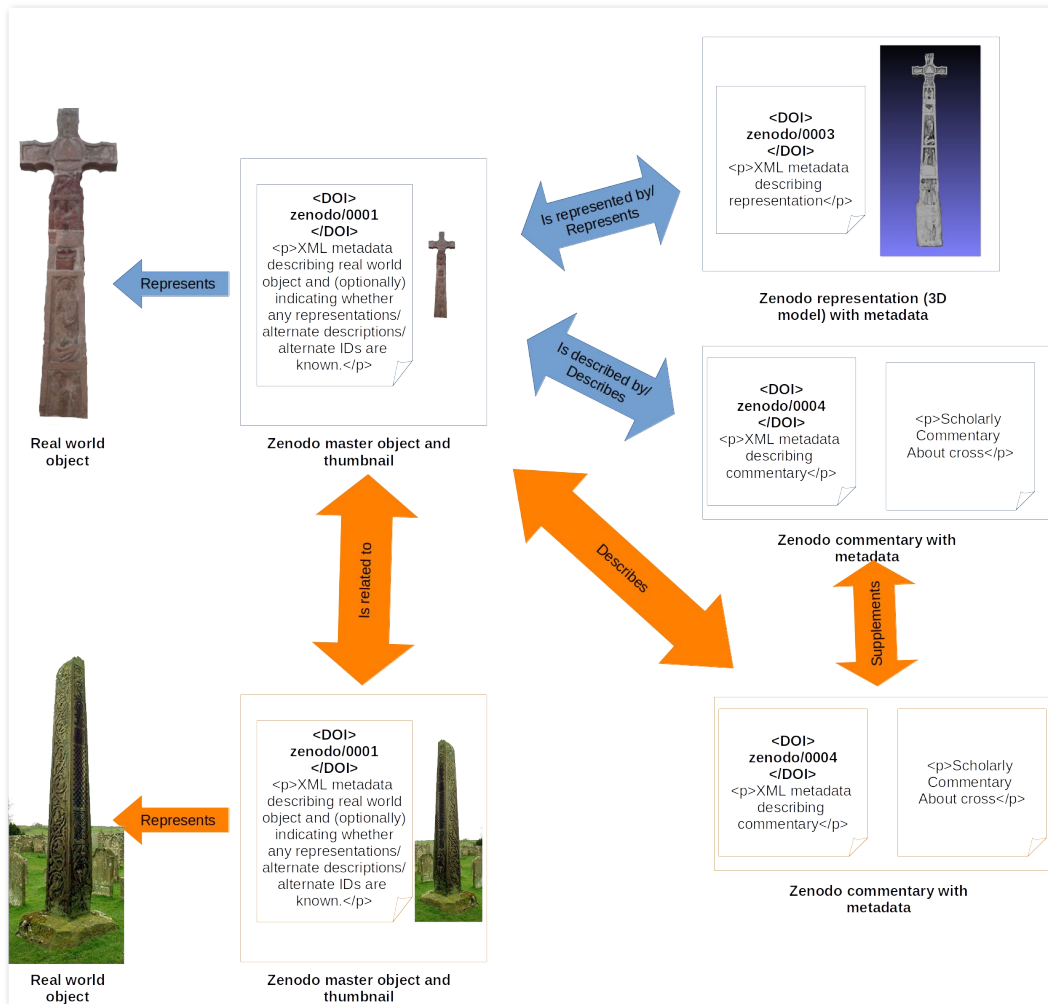


Zenodo master object and thumbnail









Advantages to this system

- Like OPenn
 - Human and Machine Readable
- Improve on OPenn
 - Persistent IDs (can be used RESTful)
 - FAIR
 - Not restricted to hierarchical arrangement or read only
 - Can be exported to variety of standards
 - Can be added to or rearranged by others
 - Maintained by archival specialists (i.e. commitment to preservation)
- **Supports small, thick, and slow publication in a FAIR format**

Disadvantages

- What is interesting about this approach is that it is accidental
 - While most features are supported,
 - Not all are (e.g. arbitrary ontologies)
 - Those that are are inconsistent across repositories (e.g. streaming; typed other identifiers)
 - Support is often tentative or inadvertent
 - Conceptual vs Record DOIs
 - Restful DOI-based API
- While the ability to support Humanities data is there, the systems have not been designed with Humanities data in mind
- Supporting small, thick, and slow data is something that can be accommodated with relatively little work

Next steps

- Next steps are to formalise this use case and feature-set
 - Build a prototype publication system within Zenodo/Github
 - Formalise and commit to the required features where they are tentative
 - Develop the few features not found specifically in Zenodo
 - Test system out on existing publications and data
 - Disseminate the model in order to encourage other systems to adopt it
- Just put together a Partnership for a SSHRC Partnership Development Grant
 - CERN/OpenAIRE
 - Toolmakers
 - Data projects
- Goal is to start prototyping this next year.

Questions