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

DOI (this version): 10.5281/zenodo.3551791
DOI (latest version): 10.5281/zenodo.3551790

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

- 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  - Thesis (1996) was analysis of (unpublished) database of textual variation in the Old English poetic canon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

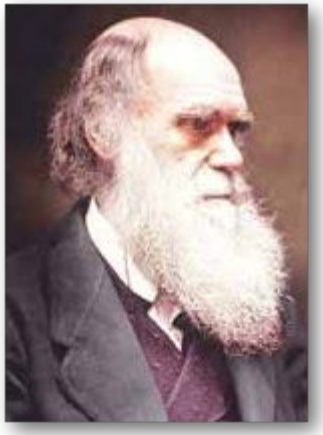
- These definitions are highly contingent
  - “Primary source” in one context, can be the “secondary source” in another (and vice versa)
  - Or simultaneously “Primary” and “Secondary” (e.g. a critical edition)
- Also hard to constrain

  “[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
- 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
- Think about Darwin and his work in the Galapagos Islands
    - What is his data?

The finches?

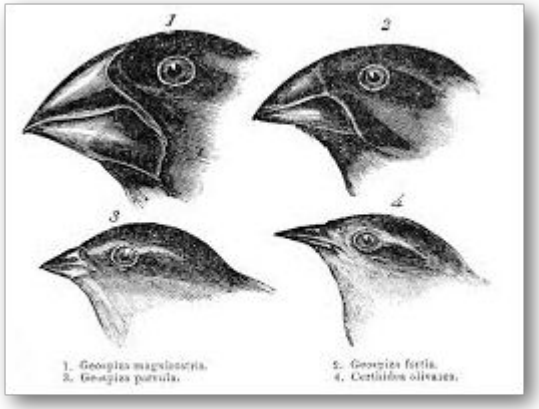
How does data work in other fields?

- 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
- 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 (later the Oxford English Dictionary) provides a non-digital model for this
  - NED was based on “historical principles” (i.e. definitions derived from and supported by historical quotations)
  - Massive crowd-sourced big-data collection effort, involving thousands of readers 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

- The solution was to create a demand and platform for new editions of medieval and renaissance works
  - Established text societies and publishers to publish new editions that met the NED’s requirements but also supported traditional humanities goals
  - Encouraged leading scholars to edit (and later reedit) the texts they needed according to the format they required
  - 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 (from both a big and a small data perspective) editions, providing the NED with the material it needed for its own big-data work

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
- In other words: work with the traditional workflows *and* do it within our systems.

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 (~~**C**~~)
    - Requires ongoing maintenance (~~**D**~~)
    - One kind of presentation (~~**A**~~)

Different approaches over the years

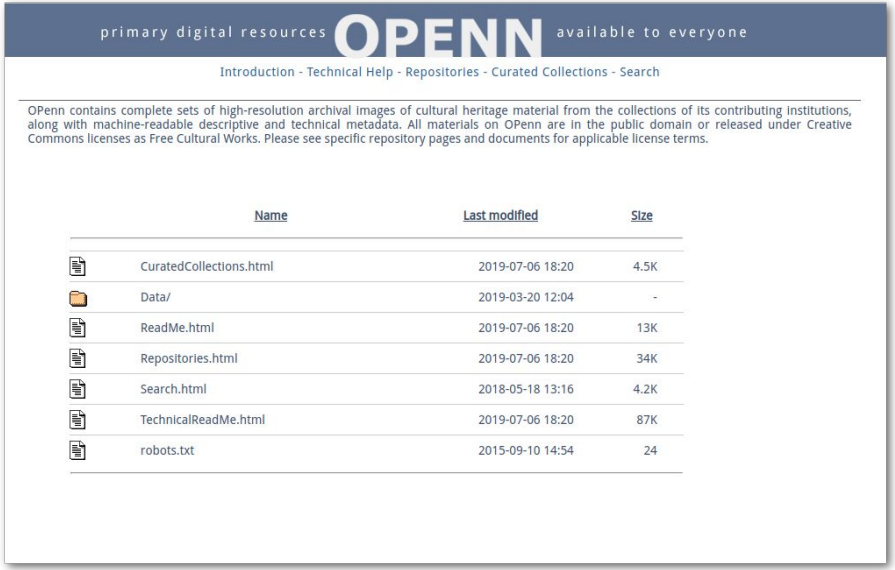
- Game engine
  - Provided different ways of organising material and good at object/collection (**A**)
  - Preserved authority (**C**)
  - Some engines allowed some external contributions (**B**)
  - But
    - Requires others to use our system (~~**B**~~)
    - None strong on external contributions (~~**B**~~)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

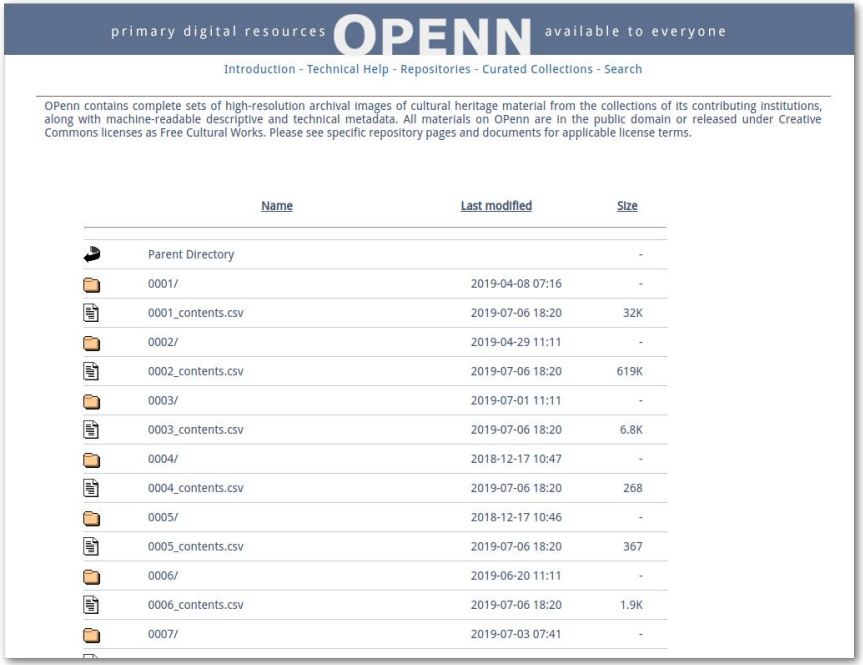


OPenn (http://openn.library.upenn.edu/)

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

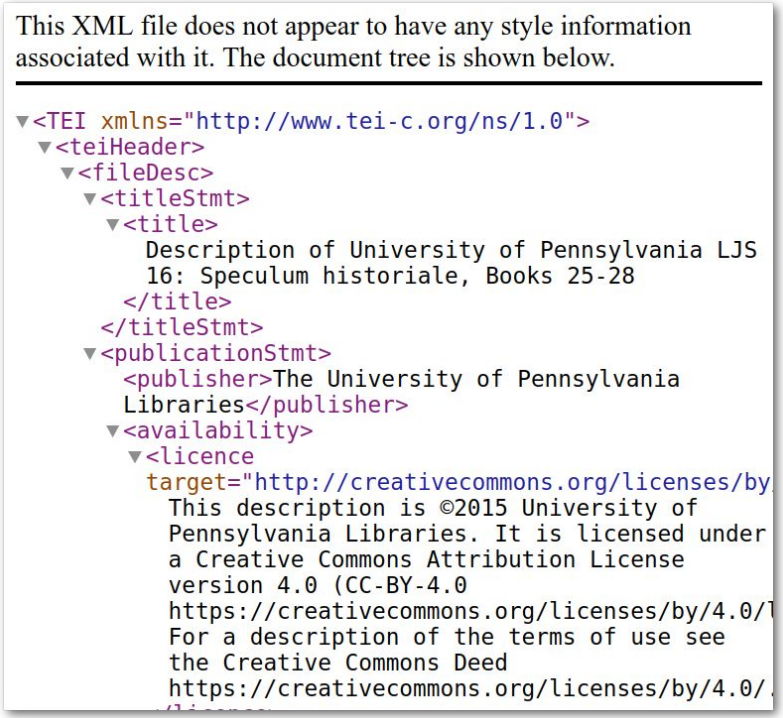
OPenn (http://openn.library.upenn.edu/)

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



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
              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/
```

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 Sienese book covers.,TEI,201
2,0001/ljs201,Evangelista Torricelli letter to Marin Marse
2014-11-03T23:38:42+00:00,2015-04-22T15:17:05+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

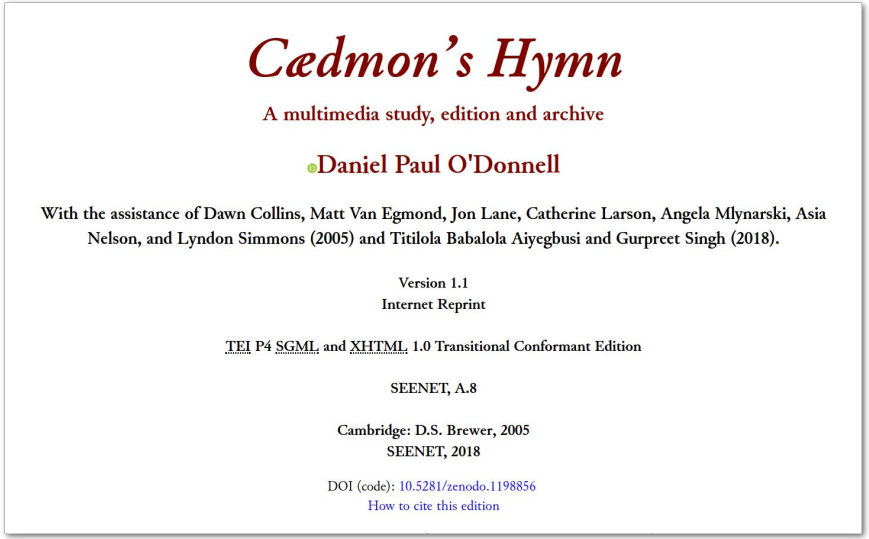
- Code repository, version control, distribution system
- Used by millions for developing code-based projects
- Recently added ability to publish web-pages using Jekyll-based “GitHub pages”
- Based on Open Source Git
- But
  - Recently bought by Microsoft (it’s always been private)
  - Not archival (conditions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

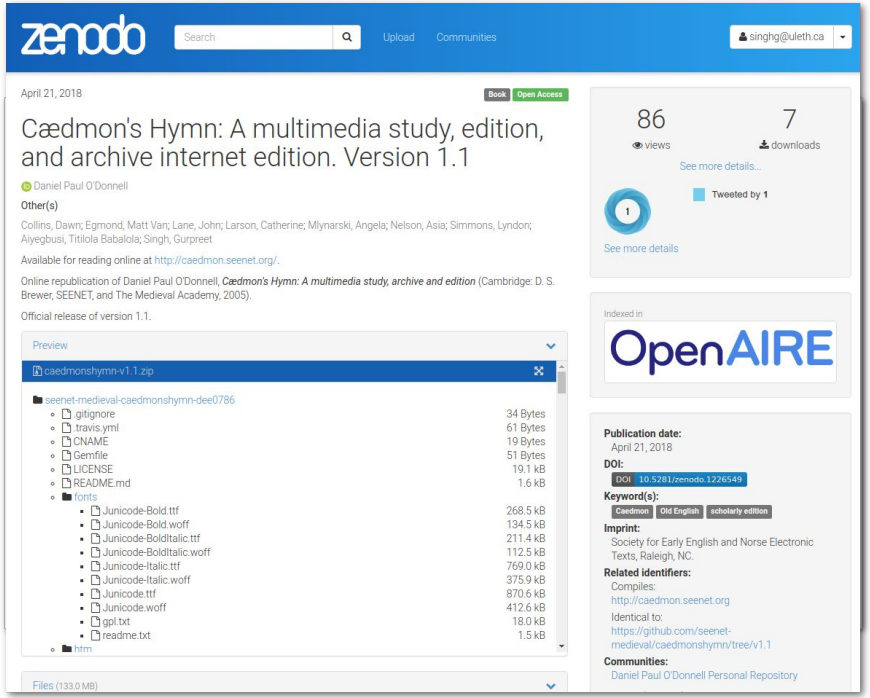
SEENET, A.8

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

DOI (code): 10.5281/zenodo.1198856
How to cite this edition

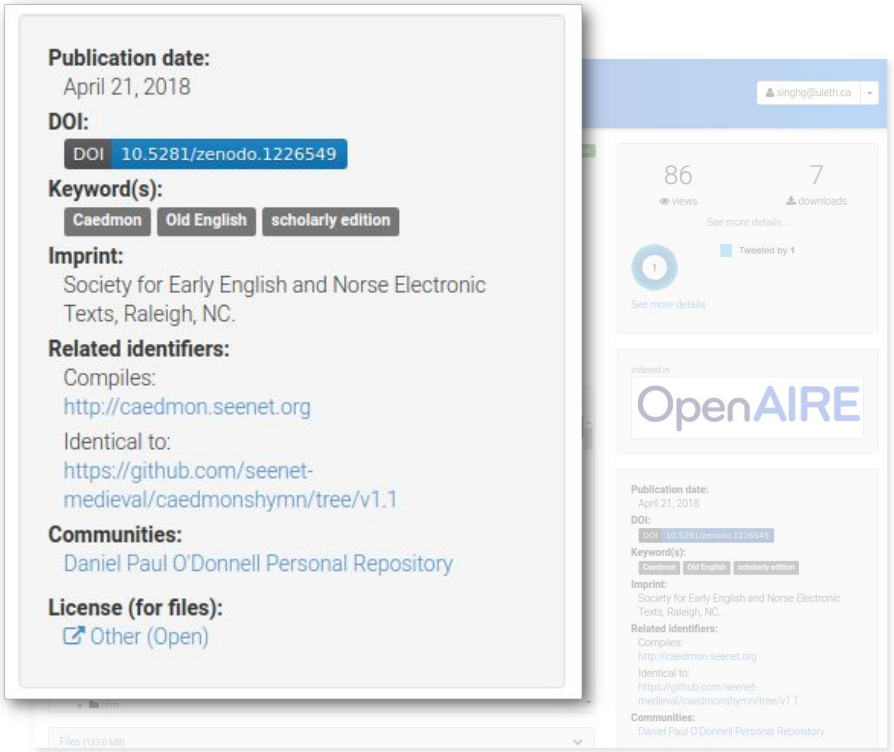
An example:
Cædmon's Hymn

- 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)



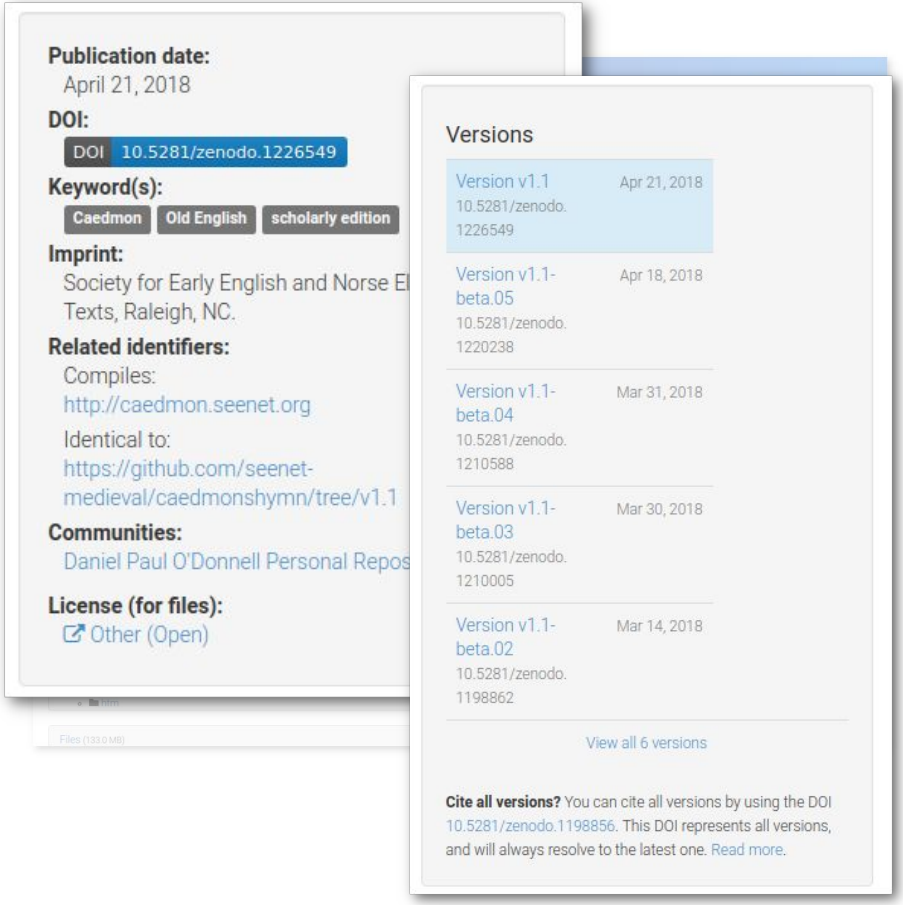
An example:
Cædmon's Hymn

- 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)



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)



Visionary Cross as Data

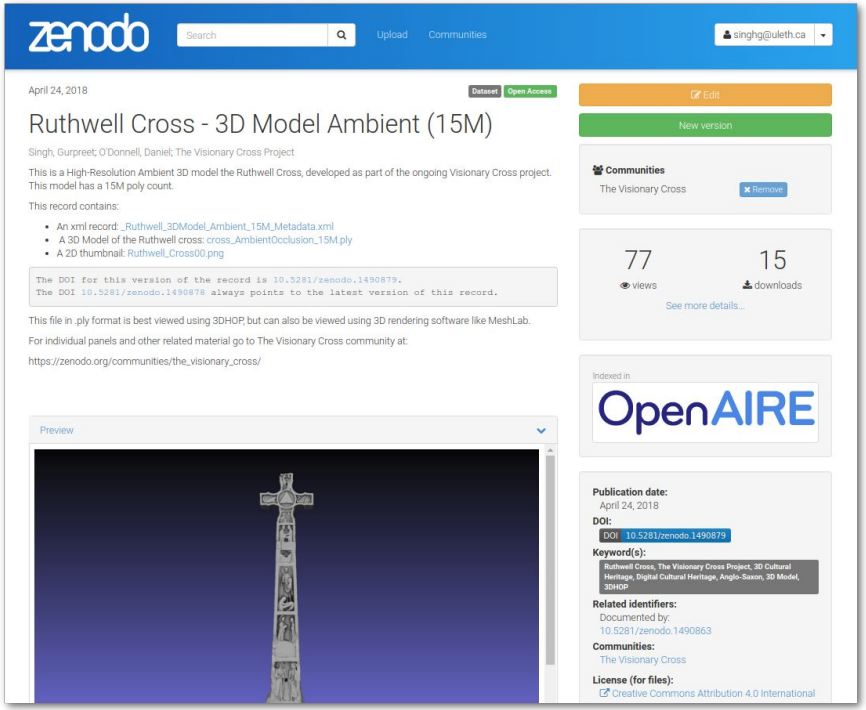
- Combining two systems allows us to publish a data-centric edition that is
  - Flexible
  - Extensible
  - Authoritative
  - Durable
  - Externally registered persistent IDs
  - Maintained by others

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

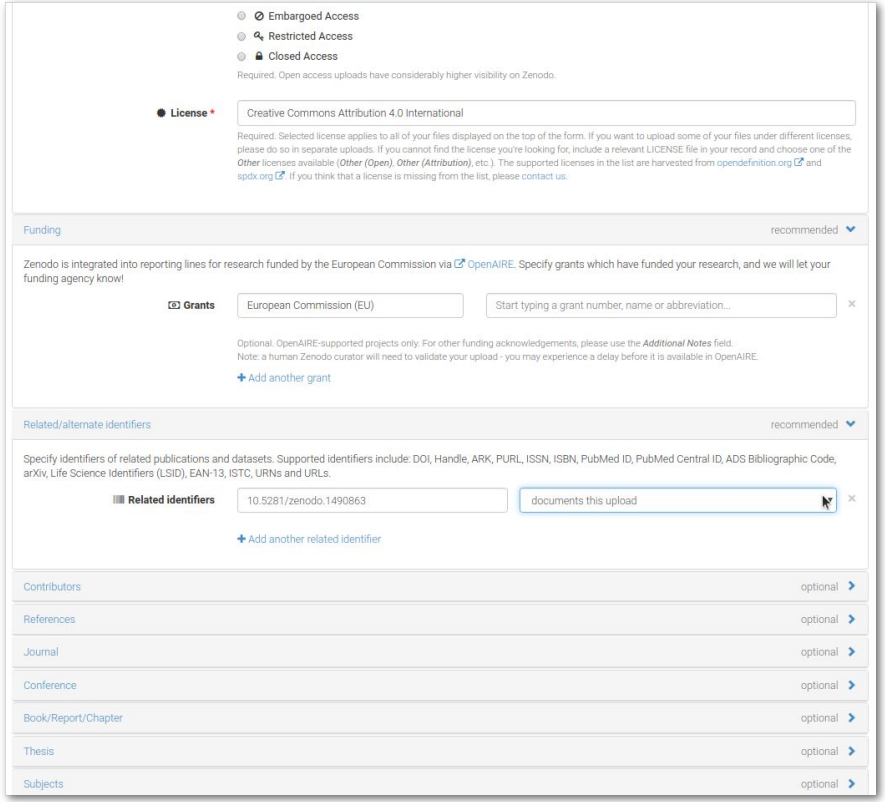
Zenodo record

- 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



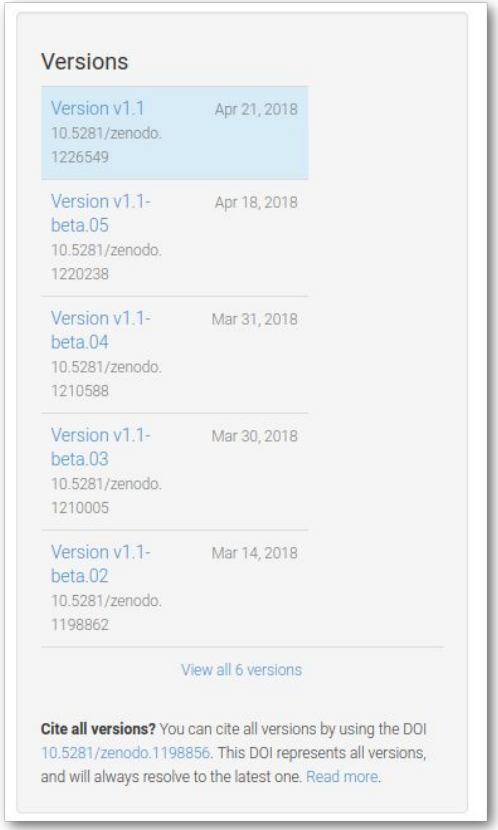
Zenodo record

- 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



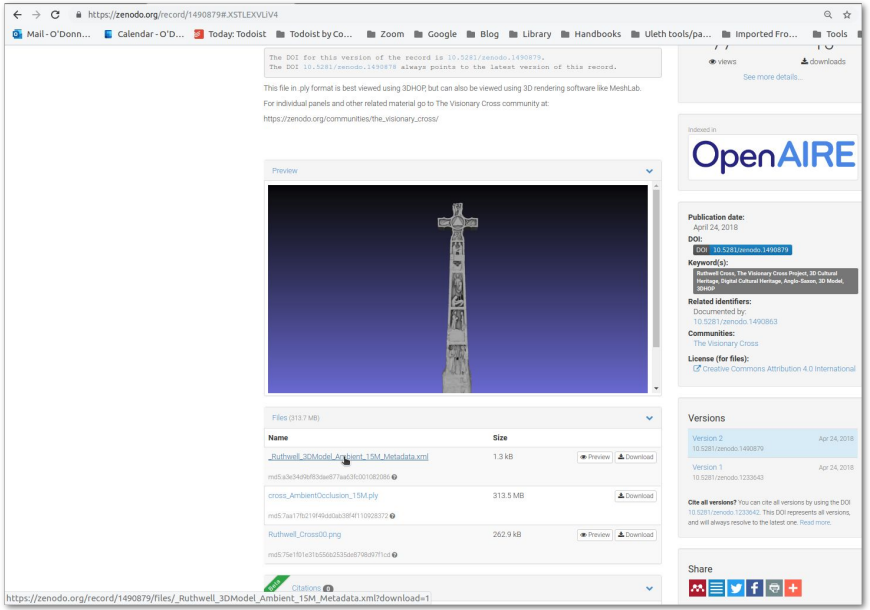
Zenodo record

- 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

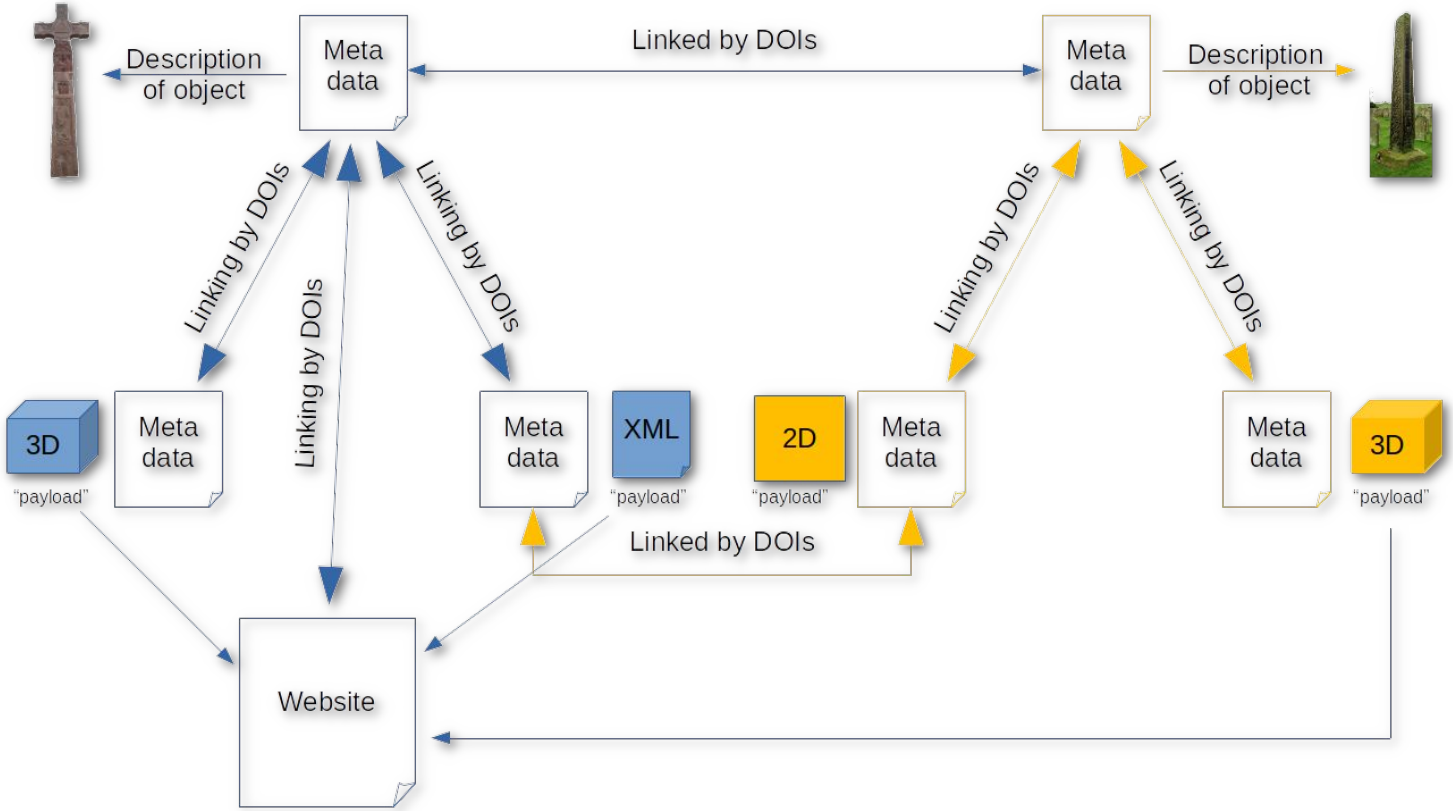


Zenodo record

- 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



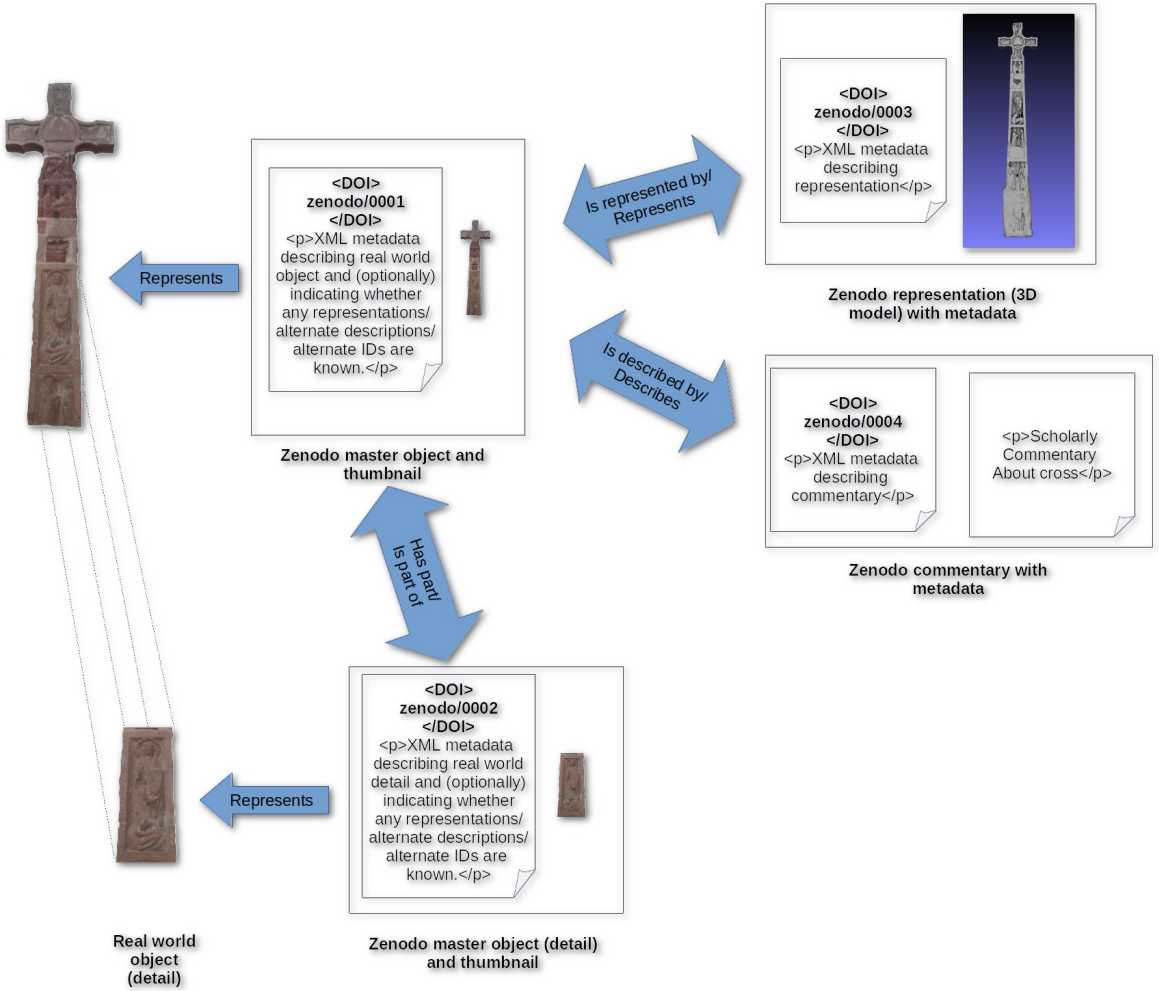
Edition is built around records

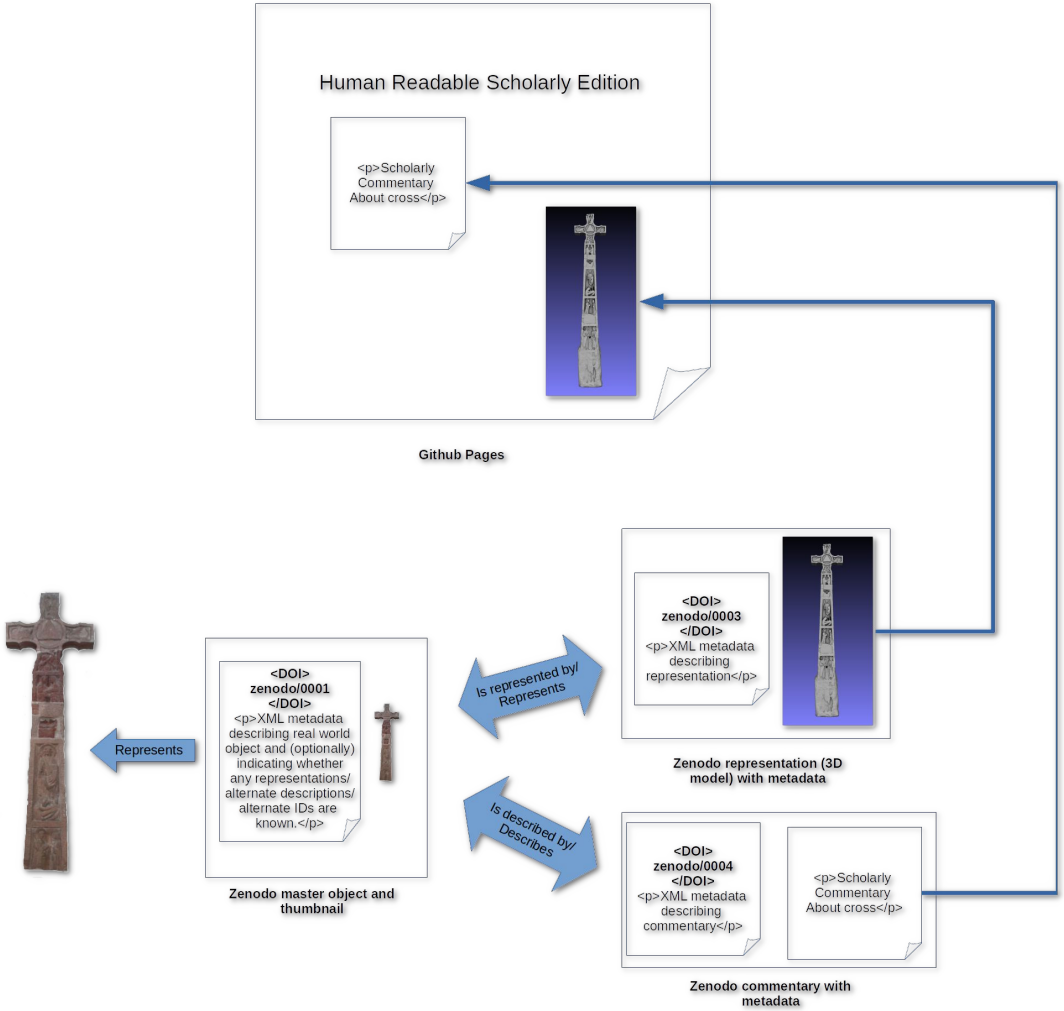


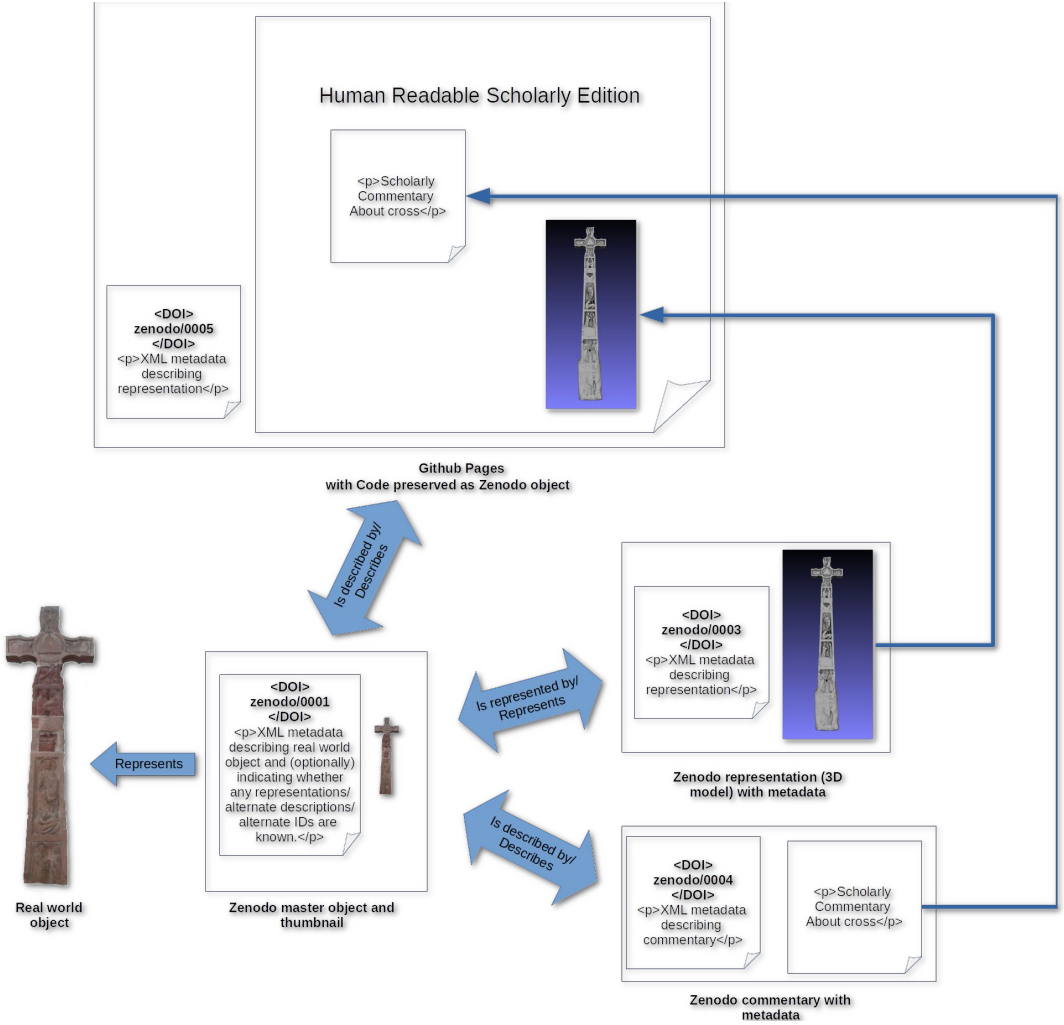
Represents

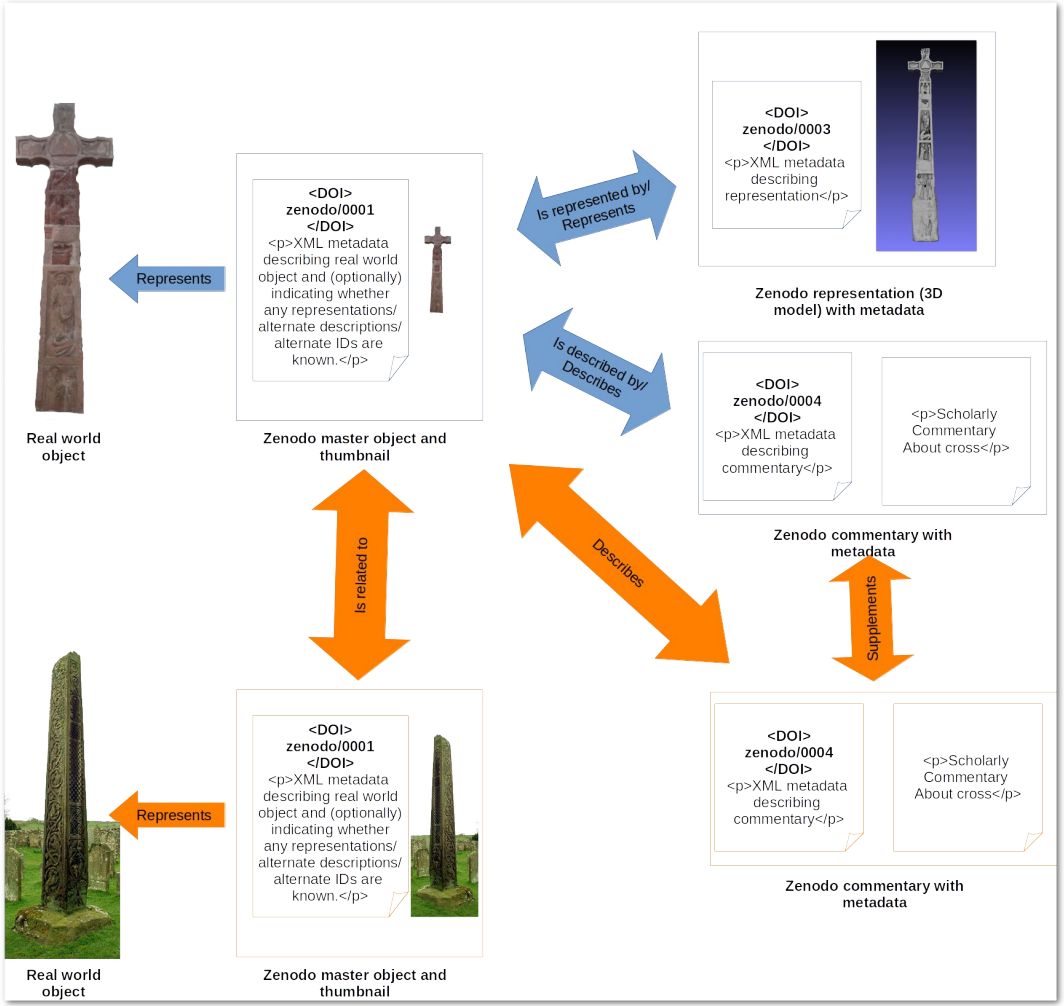
<DOI>
zenodo/0001
</DOI>
XML metadata describing real world object and (optionally) indicating whether any representations/ alternate descriptions/ alternate IDs are known.

Zenodo master object and thumbnail


<DOI>
zenodo/0003
</DOI>
XML metadata describing representation
Zenodo representation (3D model) with metadata
Is represented by/
Represents
<DOI>
zenodo/0001
</DOI>
XML metadata describing real world object and (optionally) indicating whether any representations/ alternate descriptions/ alternate IDs are known.
Represents
Zenodo master object and thumbnail
Is described by/
Describes
<DOI>
zenodo/0004
</DOI>
XML metadata describing commentary
Scholarly Commentary About cross
Zenodo commentary with metadata
Has part/
Is part of
<DOI>
zenodo/0002
</DOI>
XML metadata describing real world detail and (optionally) indicating whether any representations/ alternate descriptions/ alternate IDs are known.
Represents
Zenodo master object (detail) and thumbnail
Real world object (detail)


Human Readable Scholarly Edition
Scholarly Commentary About cross
Github Pages
<DOI> zenodo/0003 </DOI>
XML metadata describing representation
Is represented by/ Represents
Zenodo representation (3D model) with metadata
Represents
<DOI> zenodo/0001 </DOI>
XML metadata describing real world object and (optionally) indicating whether any representations/ alternate descriptions/ alternate IDs are known.
Zenodo master object and thumbnail
Is described by/ Describes
<DOI> zenodo/0004 </DOI>
XML metadata describing commentary
Scholarly Commentary About cross
Zenodo commentary with metadata


Human Readable Scholarly Edition
Scholarly Commentary About cross
<DOI> zenodo/0005 </DOI>
XML metadata describing representation
Github Pages
with Code preserved as Zenodo object
Is described by/ Describes
Is represented by/ Represents
Is described by/ Describes
Represents
<DOI> zenodo/0001 </DOI>
XML metadata describing real world object and (optionally) indicating whether any representations/ alternate descriptions/ alternate IDs are known.
Zenodo master object and thumbnail
Real world object
<DOI> zenodo/0003 </DOI>
XML metadata describing representation
Zenodo representation (3D model) with metadata
<DOI> zenodo/0004 </DOI>
XML metadata describing commentary
Scholarly Commentary About cross
Zenodo commentary with metadata


Represents
<DOI>
zenodo/0001
</DOI>
XML metadata describing real world object and (optionally) indicating whether any representations/ alternate descriptions/ alternate IDs are known.
Is represented by/ Represents
<DOI>
zenodo/0003
</DOI>
XML metadata describing representation
Zenodo representation (3D model) with metadata
Is described by/ Describes
<DOI>
zenodo/0004
</DOI>
XML metadata describing commentary
Scholarly Commentary About cross
Zenodo commentary with metadata
Real world object
Zenodo master object and thumbnail
Is related to
Describes
Supplements
Represents
<DOI>
zenodo/0001
</DOI>
XML metadata describing real world object and (optionally) indicating whether any representations/ alternate descriptions/ alternate IDs are known.
<DOI>
zenodo/0004
</DOI>
XML metadata describing commentary
Scholarly Commentary About cross
Zenodo commentary with metadata
Real world object
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