UNLOCKING DIGITAL TEXTS

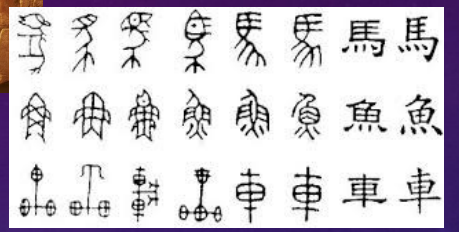
Constructing narratives across diverse corpora

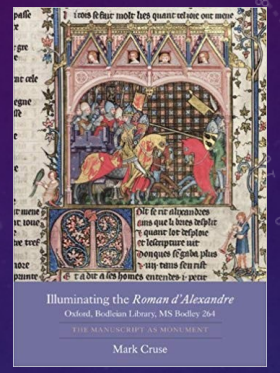
NEIL JEFFERIES, BODLEIAN LIBRARIES, UNIVERSITY OF OXFORD

DATA FUTURES GMBH

https://orcid.org/0000-0003-3311-3741

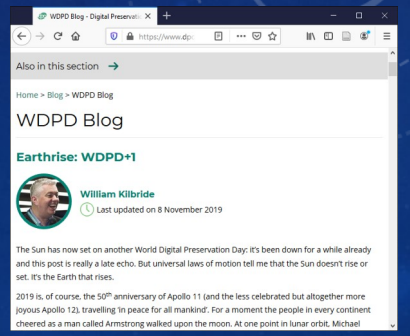
BACKGROUND





- Physical Text
  - Pictographical origins - fragmentary use
  - Developed into a serialisation of spoken language - with vary degrees of phonetic alignment
  - Longer forms emerged as a by-product of carrier technology - paper, print, post, bindings
  - Emergence of mass print led to (temporary) decrease in non-textual elements
- Digital text
  - Explosion of new (semi) pictoral forms thanks to Unicode
    - ASCII Art/smileys :-) Emoji 😀 Emoticons/Kaomoji ¯_(ツ)_/¯
  - Largely language independent but divergent culture and event specific meanings already (👌, Chinese☺)
  - Proliferation of communication channels - often used simultaneously
  - Explosion of embedded content - some encodable as text (MATHML, MEI)
- Mapping discourse and constructing narratives is hard after the event

slack



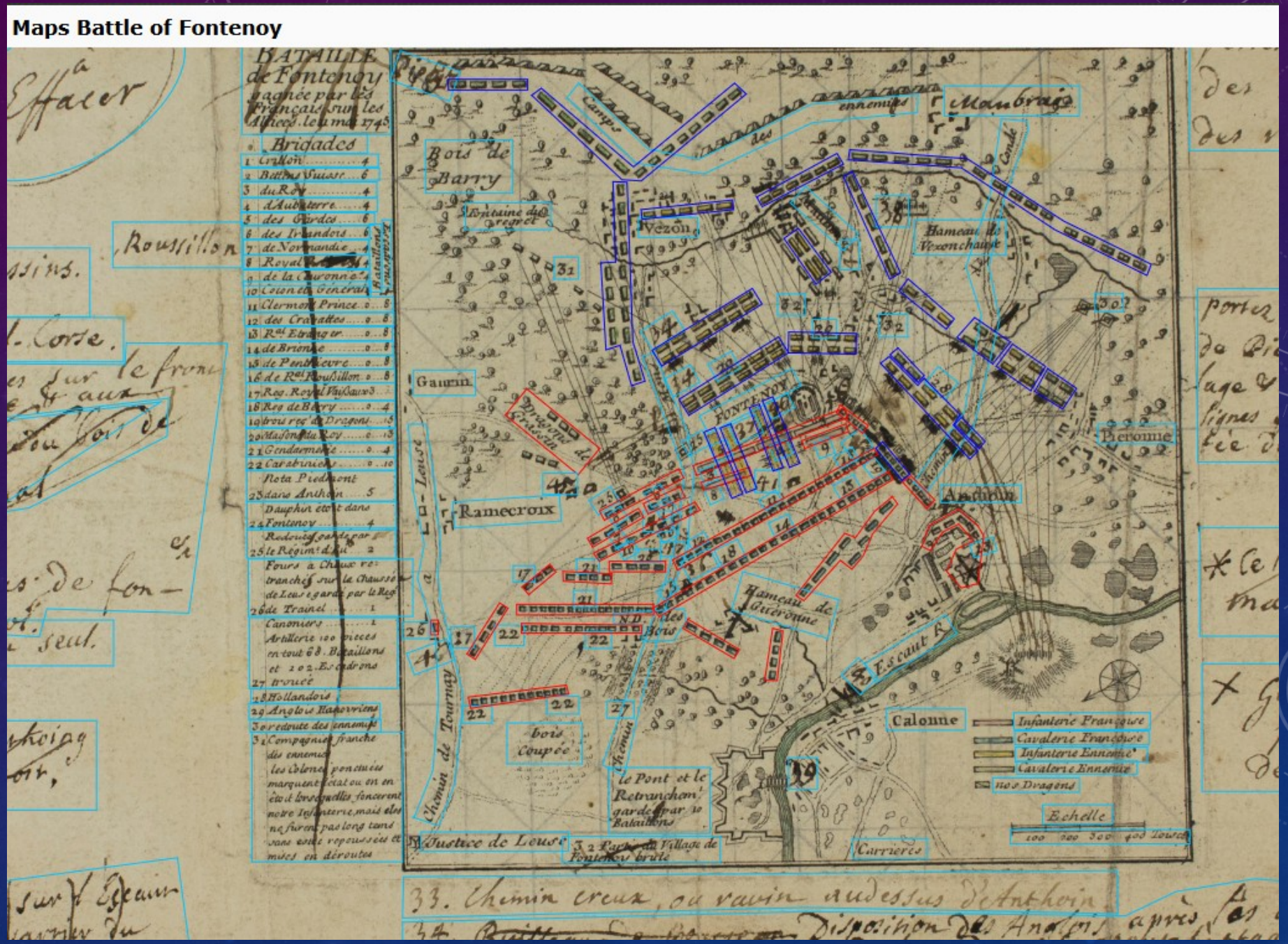
DIGITAL TEXT FORMAT DIVERSITY

- Textual materials will always be stored in variety ways to reflect diverse use-cases
  - Almost all focus on short-term utility over effective information capture (understandably!)
  - Which generally translates as: Human readability
- Document formats are mostly designed for final-format publishing
  - Ongoing use or re-use is a secondary consideration
  - Emergent formats are ill-specified, MS/Open Office Formats, PDF, SGML/XML all suffer from this
    - Implementations vary so hard to develop tools (and maintain in the long term)
    - Excessive versions, features like embedded objects and executables complicate things
- Stream formats (Twitter, email) are complex and dynamic
  - Multiple authors, versions and views
  - Navigation often requires loading into a database-like application
  - Often heavily linked to other resources, and dependent on them for meaning
- Fragmentary text has re-emerged in a big way
  - Meme captions, hashtags, tweets

THE MULTI-HIERARCHY PROBLEM

- The majority of text formats arrange text in a hierarchical manner
  - This makes a lot of sense from a presentation point of view
  - TEX, which codifies typesetting practice almost as much as it follows it, is hierarchical
    - ...and is still the only satisfactory way of handling equations (sorry, MathML)
  - XML formats such as TEI and OpenOffice, PDF, and the browser DOM follow suit
  - But even then, the format must privilege physicality or structure
    - e.g. volume-page-line vs chapter-paragraph-sentence
- When we want to talk about text, this really breaks down
  - There are many ways of approaching a text
  - Multiple hierarchies co-incident on a base text
  - Epistemologically lax – structure, semantics and interpretation mashed together
- Cramming everything into a single hierachy is hard to use/read/maintain
  - Even harder to cite/reference at a granular level

THIS->


Maps Battle of Fontenoy
BATAILLE de Fontenoy gagnée par les Français sur les Alliés le 11 mai 1745
Bois de Barry
Vezon
Hameau de Vezonchaux
Maubrais
Gaurain
Ramecroix
Fontenoy
Hameau de Gueronne
Calonne
Carrieres
Infanterie Françoise
Cavalerie Françoise
Infanterie Ennemie
Cavalerie Ennemie
nos Dragons
Echelle
Chemin de Tournay
le Pont et le Retranchem.
bois Coupée
Justice de Leuse
33. Chemin creux, ou ravin au dessus d'Antheing

STAND-OFF MARKUP/ANNOTATION BUT...

- Needs a standardised way of referencing texts and text fragments
  - Independent of underlying format
  - For citation and referencing but also actionable as a machine API
  - Requires PID's and a co-ordinate or addressing system that is higher level than just a glyph
- Separate content from rendering - especially for born digital
  - Not completely simple, line breaks matter for poems, not for essays
  - Support multiple higher-level addressing schemes based on a particular rendering/analysis of the text
- Versioning and differencing
  - Texts, especially digital ones, evolve - abstracting rendering makes differencing easier
- Linkable
  - Text fragment references peer with IIIF image/media fragments, DataCite PID's, ISNI/ORCID persons etc.
    - ...and especially embedded non-textual content

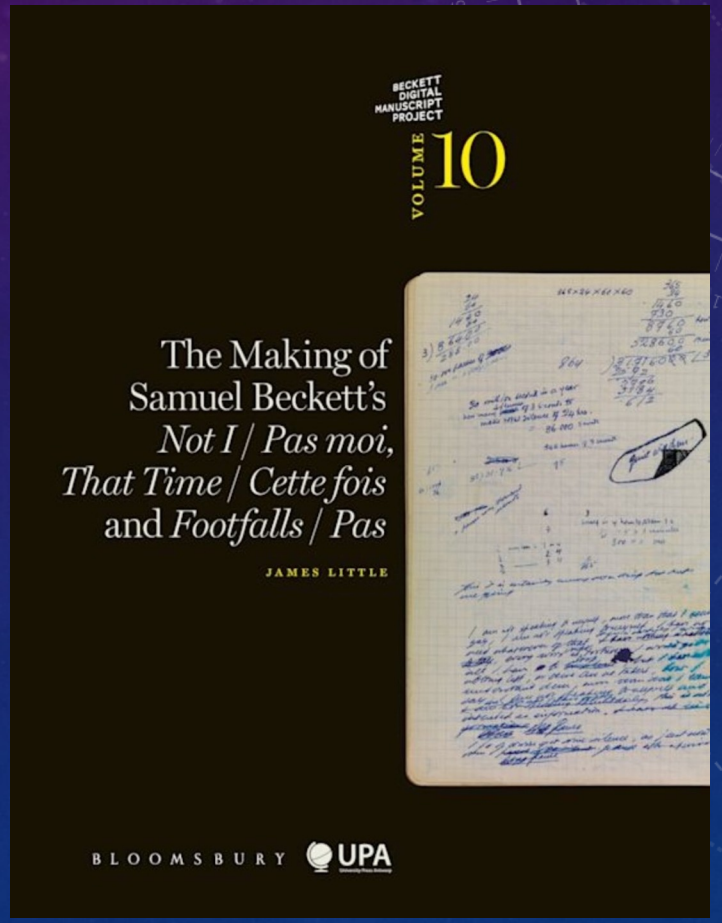
UNLOCKING DIGITAL TEXTS

- 2 year AHRC/NEH funded project “New Directions in Digital Scholarship”
  - Started 3rd May 2022, first meeting 12th
- Develop outline standards and prototypes/proofs-of-concept
  - Aim to emulate the approach used with IIIF
  - Everything will be open and available on GitHub, external contributions welcome
- Oxford, Cambridge, Notre-Dame (USA) lead partners

- Anchored in actual digital resources and the scholars that work with them
  - Sustainability/ongoing use of content
- Workshops to allow other members of the community to contribute
  - Primarily online, although in-person events are planned later in the project

OXFORD – THE SAMUEL BECKETT ARCHIVE

- https://www.beckettarchive.org/
- Mix of different document types
  - Letters, plays, Becketts personal library...
- Mix of different formats
  - TEI, PDF, digtised images...
- Mix of different approaches that span multiple documents
  - Text analytics, narratives, calendars and timelines...
- Collaboration between multiple scholars and institutions
- Currently relies on bespoke code which needs maintenance

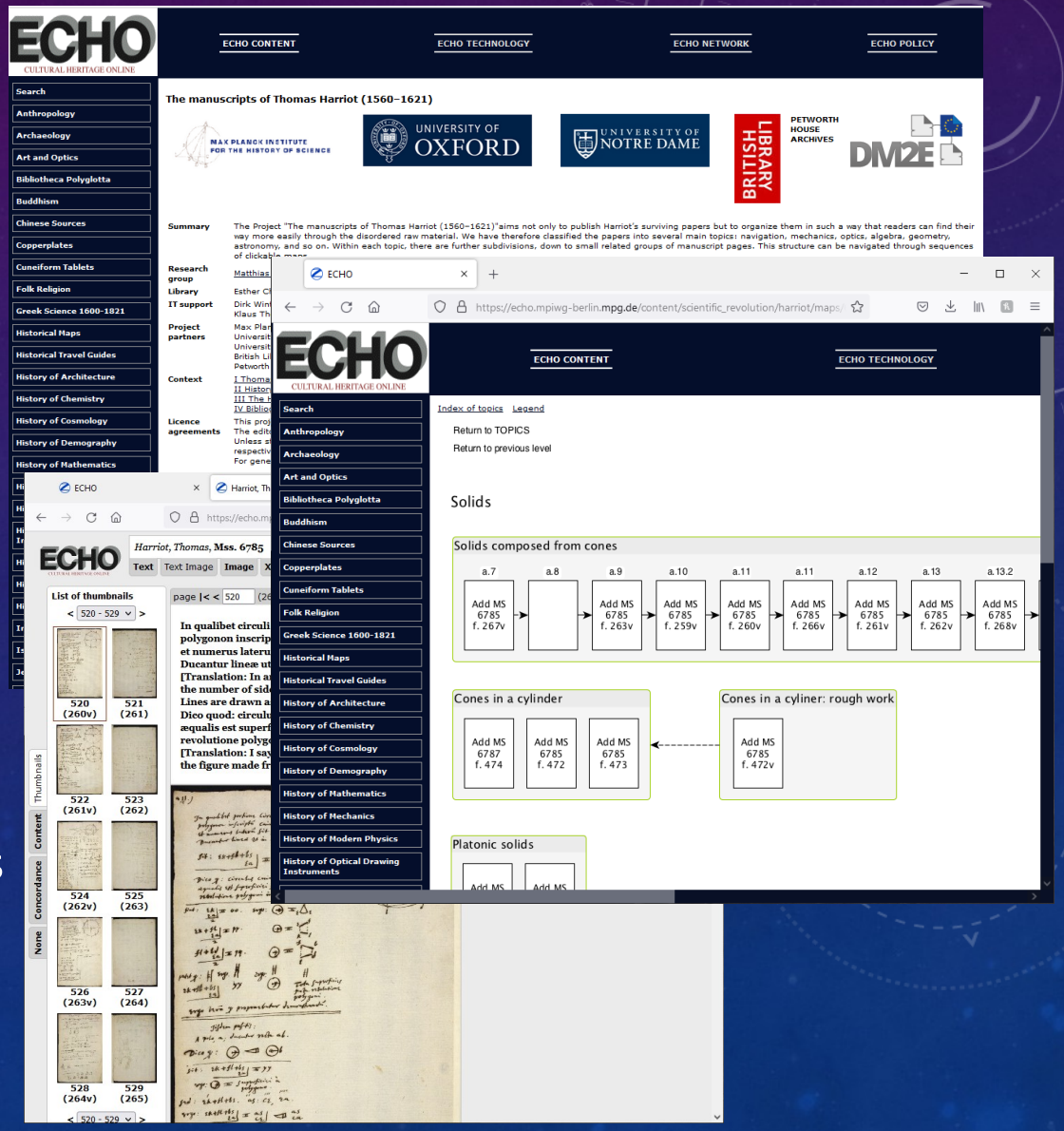


CAMBRIDGE – CASEBOOKS

- https://casebooks.lib.cam.ac.uk/
- Text fragments
  - Multiple case entries per page
- Multiple coincident narratives across scattered entries
  - Personal case histories, historical astrological and ecclesiastical analyses, modern geographical and epidemiological analyses
- Mining of TEI-XML and metadata
- Linking to other contextual resources such as Early Modern Letters Online
- Collaboration between multiple scholars and institutions
- Currently relies on bespoke code which needs maintenance

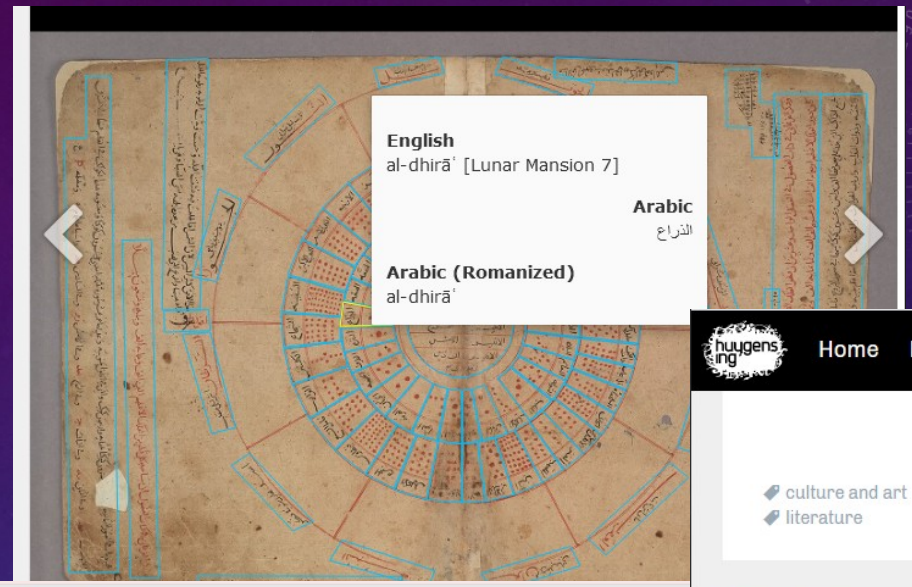
NOTRE-DAME – THOMAS HARRIOT

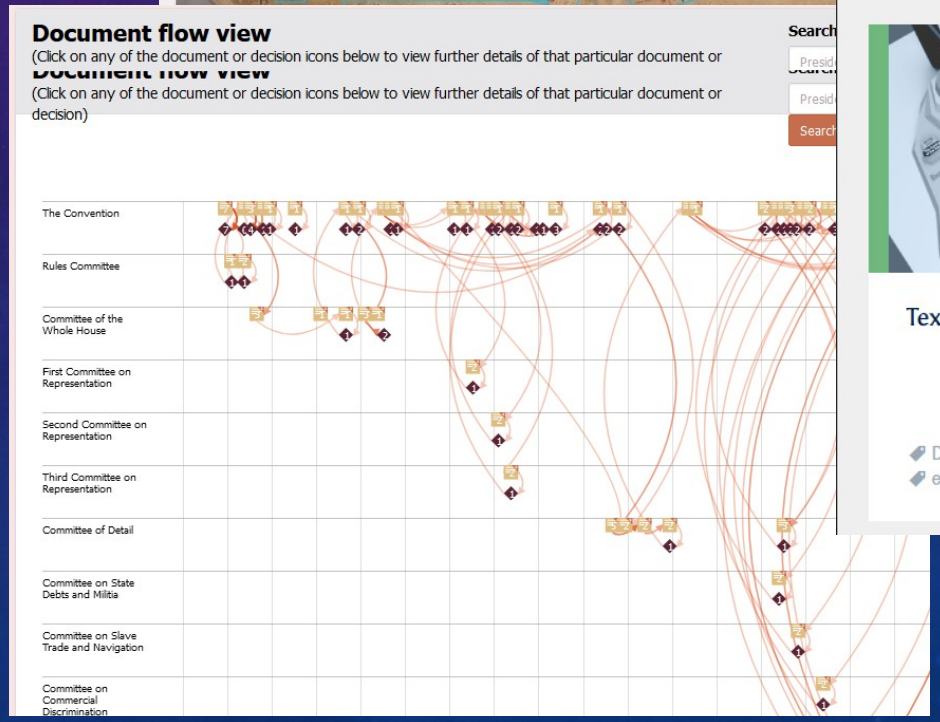
- https://echo.mpiwg-berlin.mpg.de/content/scientific_revolution/harriot
- Wide variety of formats and content
  - XML, Images/Diagrams, TEX, PDF, Harriot's cipher (based on Algonquin)
- Multiple coincident narratives across scattered entries
  - Harriot was a polymath. Any single document can reference many themes/narratives
- Material used for teaching as well as research
- Collaboration between multiple scholars and institutions
- Currently relies on very old bespoke code which needs maintenance.
  - A considerable body of existing historical analysis and work to be redelivered

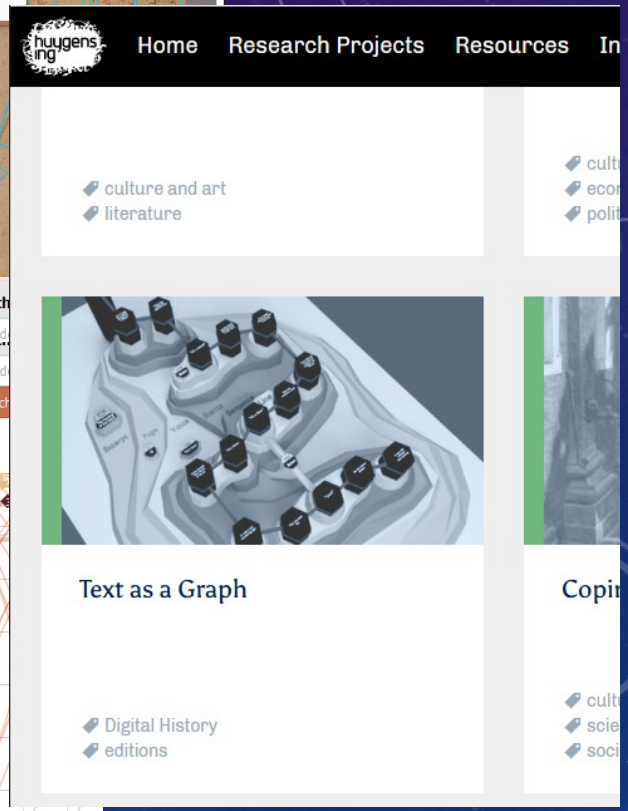


OTHERS WELCOME!

- Text-as-graph
  - Humanities Cluster, NL
- Text-as-process, negotiated texts
  - Quill Project, Oxford
- Email
  - ePADD group, Stanford, USA
- Parallel trans*tions, mixed directionality content
  - Book Of Curiosities & Digital Shikshapatri, Oxford
- ALTO
- Ideographic scripts...







WHO WE ARE (SO FAR)

- Oxford
  - Neil Jefferies, Bodleian Libraries (PI)
  - Dirk Van Hulle, Faculty of English (CoI)
  - Megan Gooch, Centre for Digital Scholarship (CoI)
- Cambridge
  - Michael Hawkins, Cambridge Digital Humanities (CoI)
  - Rob Ralley, Cambridge Digital Humanities
- Notre Dame
  - Robert Goulding, John J. Reilly Center for Science, Technology, and Values (PI)
  - Scott Weingart, Navari Family Center for Digital Scholarship (CoI)
  - Arnaud Zimmern, Navari Family Center for Digital Scholarship
  - Natalie Meyers, Lucy Family Institute for Data and Society

INSPIRATIONS

- IIIF
  - Image API provides a standardised way of accessing images and fragments largely independent of the underlying format
  - Presentation API describes how images and fragments should be assembled for consumption
  - Machine and human consumer targets
- Browser-based viewers
  - PDF, XML, Office formats are all mapped* to the HTML-DOM hierarchical object model for display
  - Hypothes.is web annotation service
- PERSEUS Canonical Text Service
  - Standard API for accessing text fragments
  - But…
    - Relies on correctly formatted text
    - Limited addressability because of focus on classical forms
    - Canonical texts rather than multiple witnesses (still final format!)
    - Database back end is inflexible and slow

EXISTING WORK

- TEI, TextGrid etc.
- IIIF Presentation API
  - Idea of assembling fragments into a presentation space
- InvenioRDM
  - Annotation Lists as a work
  - Using ORCID for authentication and attribution
- Transcribus (and others)
  - Machine generated annotations
  - Using human generated annotation to train ML systems

- API Work
  - KNAW HUC
  - Gottingen Text Mining
- Versioning
  - Memento
  - Oxford Common File Layout
- And others…
  - COST IS1310

- NO new text formats
- NO code dependencies

STRUCTURED TEXT FRAGMENT REFERENCES & API

- Texts should have unique identifiers
- Profiles for different textual forms
  - Defines addressing scheme
    - Book, chapter, paragraph
    - Anthology, Poem, Stanza etc.
  - Many similar terms for the same thing
    - Mapping (RDF vocabularies?)
  - Query to find what is supported
- Machine and human oriented
  - Bare text – Unicode, no formatting
  - TEI*
  - Text rendered for HTML display
  - Original markup

- Functional use-cases
  - Complete text retrieval
  - Fragment targets for standoff annotation markup
  - Fragment specifiers for citation/referencing
  - Access state of document at a particular time

- Application
  - Flatten source and a selected annotation/markup set into XML/TEI

STRUCTURED TEXT SERVER PROOF OF CONCEPT

- Assigns texts a persistent identifier
- Handles text versioning
- Supports a limited number of formats for fragment retrieval API
  - Raw UTF-8 Text
  - TEI*
  - TEX
  - Markdown
  - PDF*
- SGML probably not solvable

- Proof of concept, some formats left for later!
  - Office formats (usually converted to PDF in repositories)
- Generic TEI/XML problematic
  - Many variants
  - Many non-structural elements
- But, it may be possible to parse files and generate…
  - TEI containing only text structure elements
  - Standoff annotations for different classes of non-structure elements
    - Which reference the base TEI via the API

Text Interoperability Framework - Preparation

Starting Point

PDF of a published paper or edition plus associated bibliographic metadata in a repository

PDF Extraction

- Structured text using a constrained TEI-derivative that defines a "text-coordinate" system for locating other entities such as...
  - Non-textual figures and images as distinct objects
  - Links (references, citations etc.)
  - PDF-specific rendering details
- Additional metadata extracted from PDF headers

Layering

- Extracted entities are layered over the structured text using annotations which can be easily rendered via HTML or more interactively using IIIF viewers
  - Images and figures (zoomable)
  - Links (possibly with some resolution)
- PDF rendering details could be used to provide CSS hints

Enrichment

- Text analytics provides scope for enrichment creating addtional annotation layers and enhanced discovery metadata
  - Named entity extraction - person, place authorities
  - Time reference extraction - calendars and timelines
  - Topic/lexical analysis - ontologies and taxonomies
- Feature extraction on images can also be applied

Text Interoperability Framework - In Use

Multiple Lenses

- Each annotation layer provides a "lens" through which the text and embedded non-textual elements can be viewed.
- Layers are orthogonal and intersect through the underlying text.
- Each layer defines ways of referencing the text at higher levels of abstraction.
- Human or machine generated annotations with provenance.
- Annotation mechanism is extensible
- Annotations reuse ontologies and taxonomies:
  - ALTO for layout
  - Specialist TEI
  - PROV-O/SEM for events
- Solves several key issues:
  - XML multiple hierarchy problem
  - The multiplicity of TEI variants
  - Difficulty of reusing content

Multiple Narratives

Annotation allows the construction of multiple scholarly narratives different approaches or opinions, but cross referencing around base text.

- Historical: people, place and event references.
- Linguistic: language structure and style
- Critical: topics. references and terminology

Multiple Texts

Narratives and annotations can reference multiple texts and use common "lenses" as a mechanism for identifying correpondences between them.

- Tracing the evolution of ideas, langauage or text through mutiple versions
- Mapping transcriptions, transliterations and translations to one another so that annotations on one carry through

UNLOCKING DIGITAL TEXTS

Constructing narratives across diverse corpora

NEIL JEFFERIES, BODLEIAN LIBRARIES, UNIVERSITY OF OXFORD

DATA FUTURES GMBH

https://orcid.org/0000-0003-3311-3741

LINKS

- TEI - Text Encoding Initiative
- TEI/XML Multiple Hierarchy Problem
- IIIF - International Image Interoperability Framework
- OCFL - Storing Versioned Digital Objects on a filesystem
- Reassembling the Republic of Letters - Contextual Models for distributed corpora
- Emoji and Emoticons
  - Divergent emoji meanings
- TextGrid, Gottingen - Looking at cross mining divergent text formats
- Humanities Cluster, KNAW - Text-as-graph but also analytics and contextuals models
- RDA WG - Preserving Scientific Annotation
- Transkribus - IIIF Human transcription Annotations used to train ML to make further Annotations