

IIIF Conference 2022

From text to image

Linking TEI-XML and IIIF via expert sourced annotations and IIIF Change Discovery

Contents

Overview

Role of IIIF

Discovery interface

TEI-XML

Madoc

Change Discovery and
Synchronisation

Use Case

The project is a scholarly project based, initially, around a single Zoroastrian religious text.

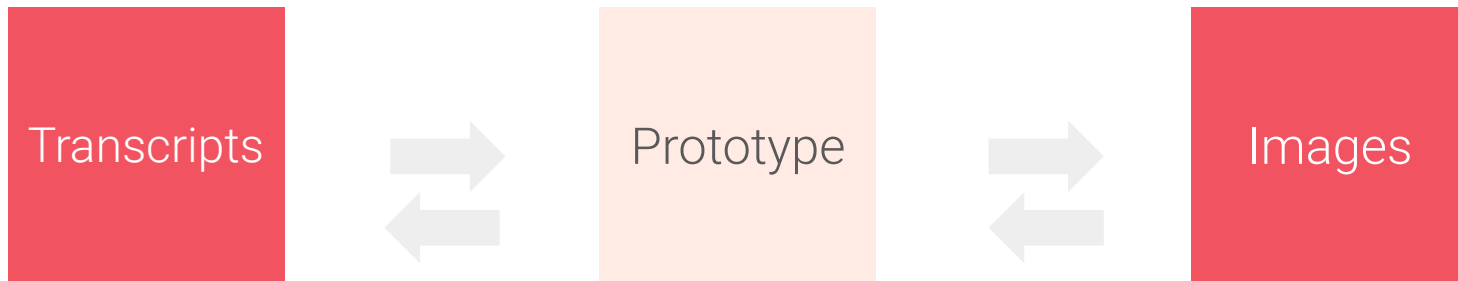
This text has multiple manuscript witnesses in many libraries and institutions, both large and small.

The project is a prototype to explore approaches to bringing together transcripts and digitised images in a single environment that allows users to explore and compare multiple instances of the same text.

This presentation is not the final prototype for that project. Rather, it's a look at the kinds of IIIF centred workflows and processes that were used to realise that prototype.

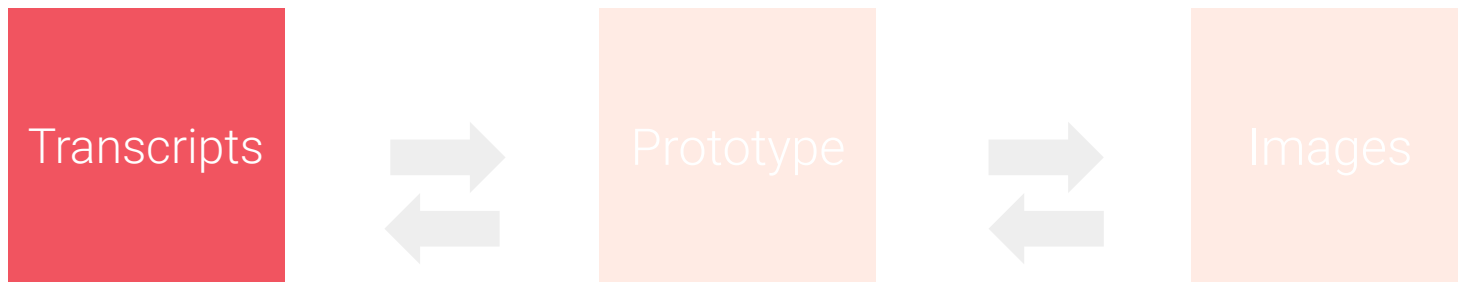
Two Sources of Information

- Transcripts
- Digitised manuscript images



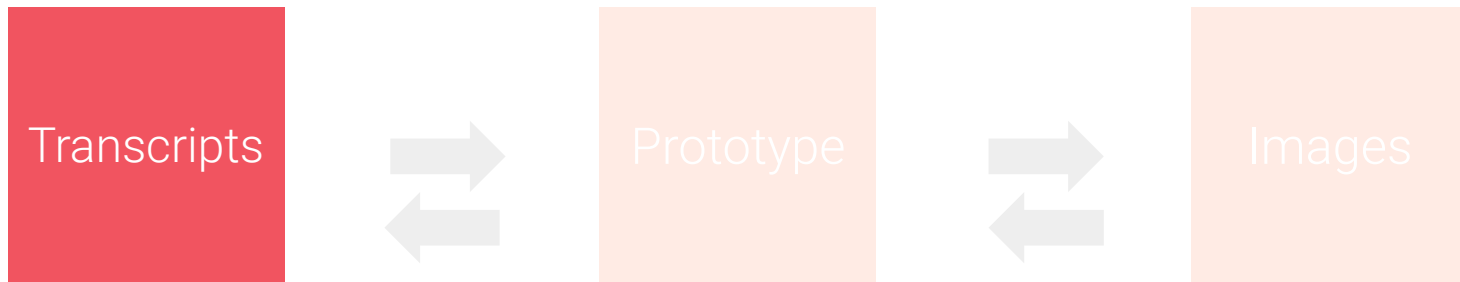
Transcripts

- TEI-XML
- Already exist or in the process of being created



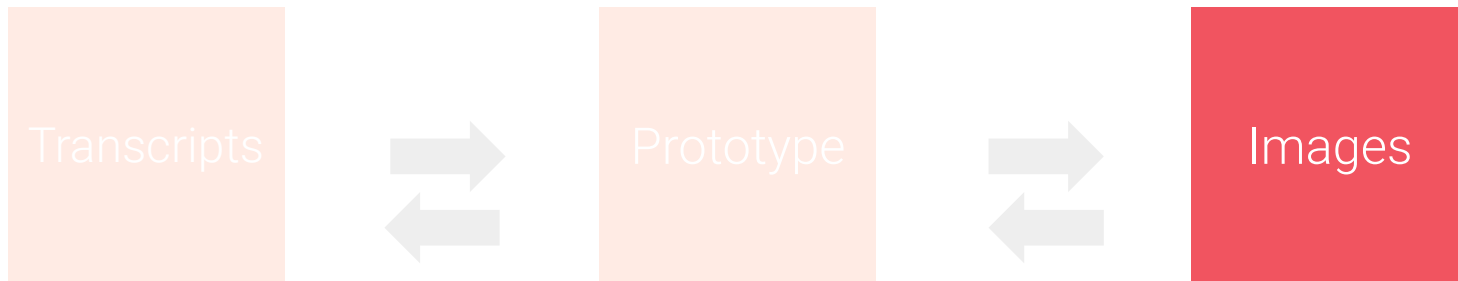
Transcripts

- No possibility to change the format or structure of the data
- TEI-XML well understood by the domain experts
- TEI is the format of choice for this project



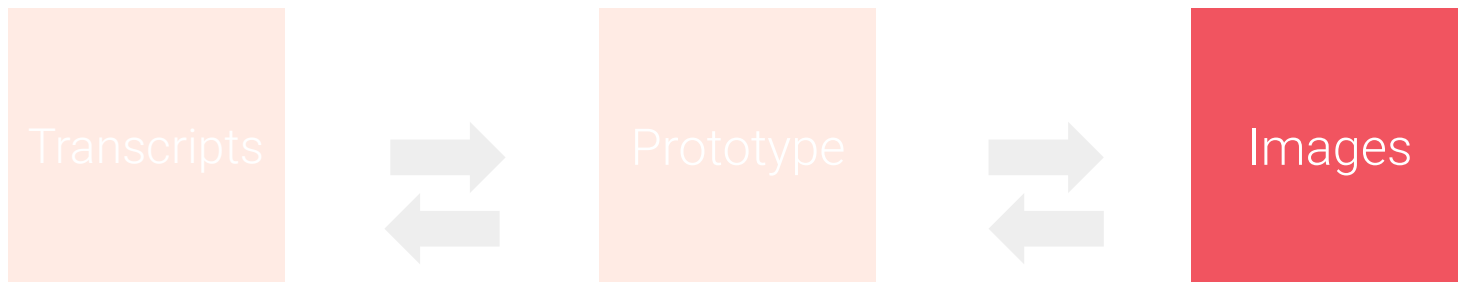
Images

- Already some existing IIIF at
 - British Library
 - Bodleian
 - KB in Copenhagen and others

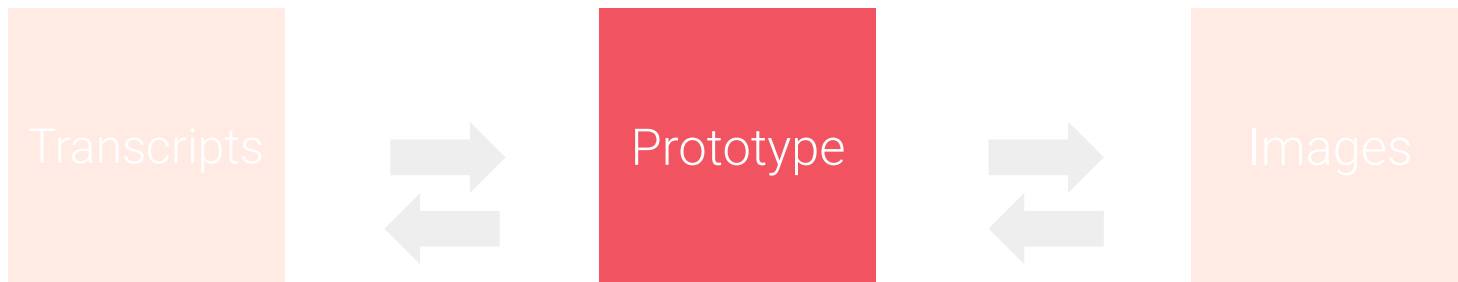


Images

- Opportunity to create IIIF from digitisation where IIIF doesn't already exist
- IIIF is the obvious format of choice for this type of resource

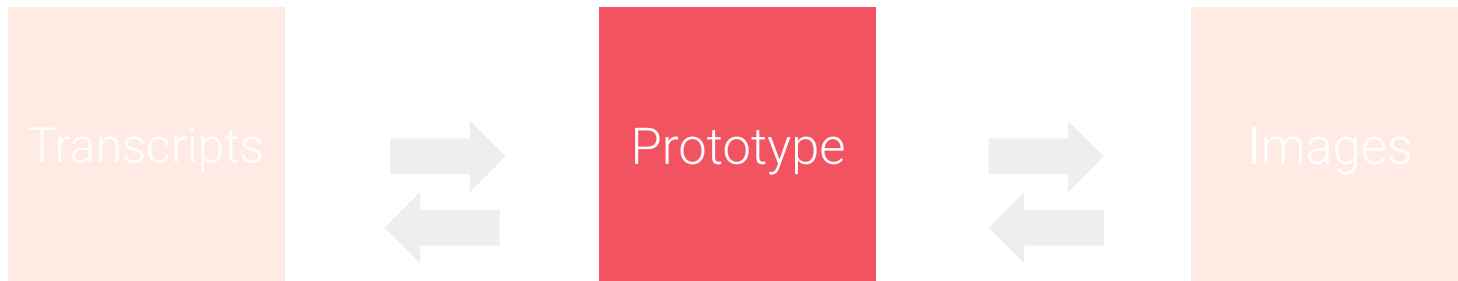


What should we build?



Requirements

- Image viewing
- Side by side image comparison
- Combination of images and text



**What we are describing here are core
IIIF use cases going back to the
earliest drafts of the IIIF Image API
and IIIF Presentation API**

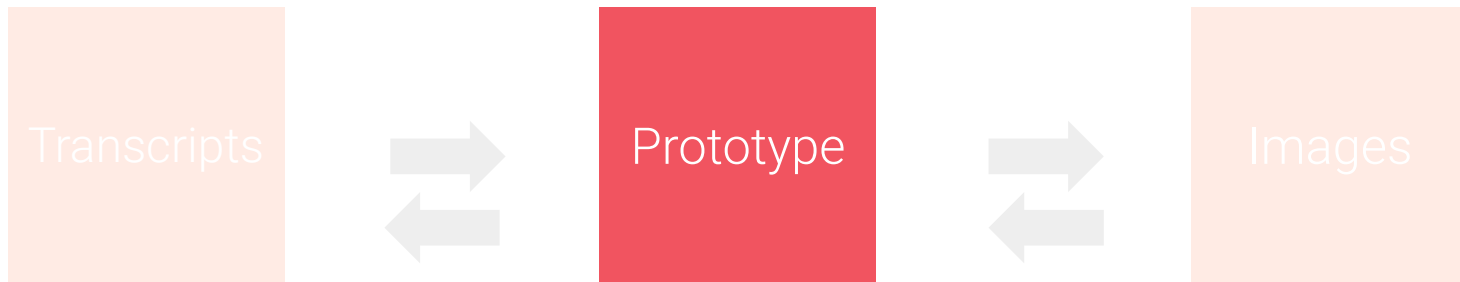
So can we just create some IIF manifests, throw them into Mirador or the Universal Viewer, and job done?

It's not quite as simple as that.

Process and workflow matters as much as data.

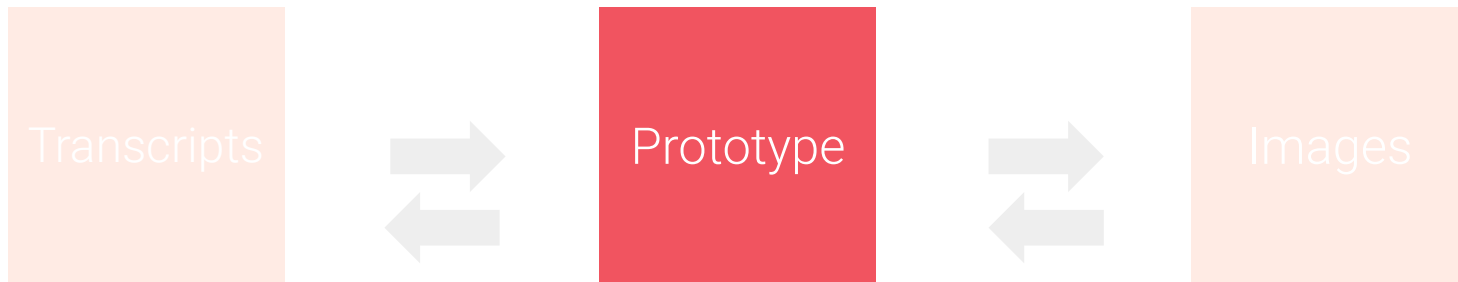
Requirements

- How do we bring together the pre-existing TEI and IIIF?
- How do we make use of document structure to navigate IIIF resources?



IIIF

- How do we model the data in IIIF?
- How do we leverage IIIF in our workflows?



IIIF as data model

As source:

- [Ranges/Structures](#) to provide navigable elements corresponding to the structure of the text
- [Annotations](#) and/or [seeAlsos](#) to link text/data with Canvases

IIIF as process enabler

As source:

- [IIIF Change Discovery](#) to synchronise between environments
- [IIIF Content State](#) for a consist way of invoking the viewer / discovery environment with specific manifests and canvases visible

Role of TEI-XML

As source:

- [TEI-XML](#) can provide transcript text for the structural elements within the manuscript(s)
- [TEI-XML](#) can provide the structure of the document as:
 - Book
 - Verse
 - Chapter

Bringing TEI and IIIF together

How?

- “Expert-sourcing” annotations in Madoc (crowdsourcing environment)
- No need to re-transcribe the data, instead, users can **link** IIIF elements with TEI-XML elements via a simple autocomplete interface, which provides (from the TEI XML) ids for:
 - Book
 - Verse
 - Chapter

What did we actually build?

An all new Discovery environment:

- IIIF Change Discovery for data import and synchronisation
- IIIF Content State for viewer state and reusable URIs
- Support for comparison
- Agnostic about metadata
- Flexible search

TEI Storage and transformation

- TEI to HTML for display
- TEI to JSON for autocomplete APIs

Madoc enhancements

- Support for TEI JSON service for autocomplete
- Support for IIIF Change Discovery for data export

Discovery UI

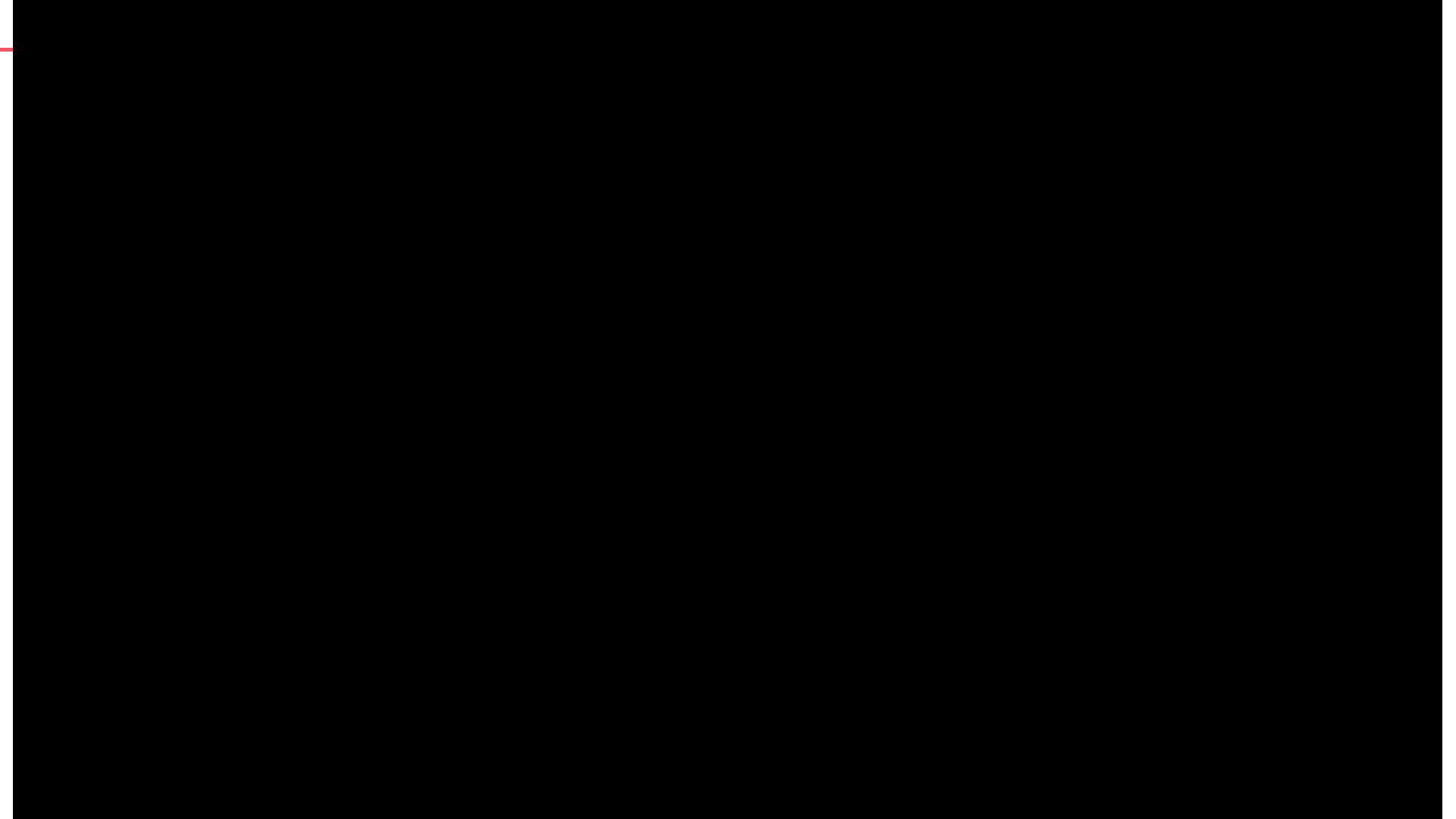


Discovery UI: Features

- IIIF Change Discovery for data import and synchronisation
- IIIF Content State for viewer state and reusable URIs
- Support for comparison via Mirador as the embedded viewer
- Agnostic about metadata
- Flexible search

Quick Tour

<https://bit.ly/dig-ch-discovery>



Book

Select book

Chapter

Select chapter

Verse

Select verse

Clear filters

Refine search

Author (10 options)

Catalogue identifier (10 options)

Collection (10 options)

Date statement (10 options)

Holding institution (10 options)

Language (10 options)

Material (1 option)

Materials (10 options)

Place of origin (10 options)

Provenance (10 options)

Publisher (5 options)

Shelfmark (10 options)

Enter keywords

Search

Showing 1 to 25 of 813 results

1

2

...

33



Bodleian Library MS. Thurston 11

Manifest:

Hits:

Metadata:

Author: Euclid, Hypsicles, of Alexandria

Pin item

View item



Bodleian Library MS. Fairfax 16

Manifest:

Hits:

Metadata:

Author: Chaucer, Geoffrey, -1400, Clanvowe, John, Sir, 1341?-1391 (?), Charles, d'Orléans, 1394-1465, Lydgate, John, 1370?-1451?, Hoccleve, Thomas, 1370?-1450?

Binding: Rebound, 2015-16, by Arthur Green and Sabina Pugh at the Bodleian conservation workshop. The boards from the former binding are housed with the volume. Light brown split calf over laminated pulpboard, 19th century; repaired 1949.

Pin item

View item

Discovery UI: Change Discovery

- Can [register](#) one or more streams to be checked at configurable intervals
- For the demo, we imported some content from the Bodleian's Change Discovery feed which is registered with the IIIF Registry.
- The site also published Change Discovery feeds so newly added annotated content can be discovered and loaded elsewhere

Django administration

WELCOME, DEMO_ADMIN [VIEW SITE](#) / [CHANGE PASSWORD](#) / [LOG OUT](#)[Home](#) > [liif_Sync](#) > [Change discovery streams](#) > [https://iiif.bodleian.ox.ac.uk/iiif/activity/all-changes](#) - 1440 min (from 2022-06-01 16:52:23+00:00)

Start typing to filter...

AUTHENTICATION AND AUTHORIZATION

Groups [+ Add](#)**Users** [+ Add](#)

DJANGO Q

Failed tasks**Queued tasks****Scheduled tasks** [+ Add](#)**Successful tasks**

«

IIIF_SYNC

Change discovery streams [+ Add](#)

YASNA

Yasna object images [+ Add](#)

Change change discovery stream

[HISTORY](#)**[https://iiif.bodleian.ox.ac.uk/iiif/activity/all-changes](#) - 1440 min (from 2022-06-01 16:52:23+00:00)****Change Discovery endpoint:****Currently:** [https://iiif.bodleian.ox.ac.uk/iiif/activity/all-changes](#)**Change:** **Sync frequency mins:****Sync from:****Date:**

Today |

Time:

Now |

Note: You are 1 hour ahead of server time.

Schedule: [+](#)[Delete](#)[Save and add another](#)[Save and continue editing](#)[SAVE](#)

TEI-XML as APIs



How did we use TEI-XML in the project? We were not interested in building tooling for creating and editing TEI. The project, which had been running before we joined, already had good tooling for TEI editing and storage.

We were interested instead in how we could use TEI **as part of our annotation, indexing and discovery workflows.**

Use of TEI-XML

As source:

- TEI-XML can provide transcript text for the structural elements within the manuscript(s)
- TEI-XML can provide the structure of the document as:
 - Book
 - Verse
 - Chapter

As an API:

- Textual elements from TEI as JSON/HTML for indexing into search and rendering in the Discovery UI
- Structural elements from the TEI-XML as an autocomplete endpoint that can be used in tagging the resources in Madoc

What did we actually build for TEI?

TEI Storage API

- Stores entire TEI-XML documents with a simple REST API
- Recursive TEI Parser which automatically breaks apart TEI-XML into fragments which can carry text and which are identified with an id
- Rendering of TEI-XML fragment content as JSON and as HTML for reuse by other APIs and for display alongside IIIF

TEI Autocomplete API

- Provides a list of endpoints (one per TEI-XML document) that can be consumed by tagging applications
- Provides autocomplete by identifier to return a list of matching identifiers by:
 - Book
 - Chapter
 - Verse
- Autocomplete in a format understood by Madoc for crowdsourcing

TEI Related Demo links

[Admin endpoint for Discovery](#)

[TEI Autocomplete list for
Discovery](#)

[TEI Autocomplete with a query](#)

[TEi Document Endpoint](#)

Caveats

Opinionated APIs:

- TEI is a very flexible and expansive format
- We did not build a universal TEI parser that could parse any incoming TEI-XML into API data
- Instead, we based our expectations around the existing identifier schemes in use on the project (which are shared across multiple institutions and projects) and the existing TEI encoding scheme

However:

- It is not that hard to customise the parser if required
- For this demo, I generated a skeleton TEI-XML file with identifiers in the expected form for The Bible without any difficulty
- Future work can expand and make more flexible the TEI parsing and TEI APIs

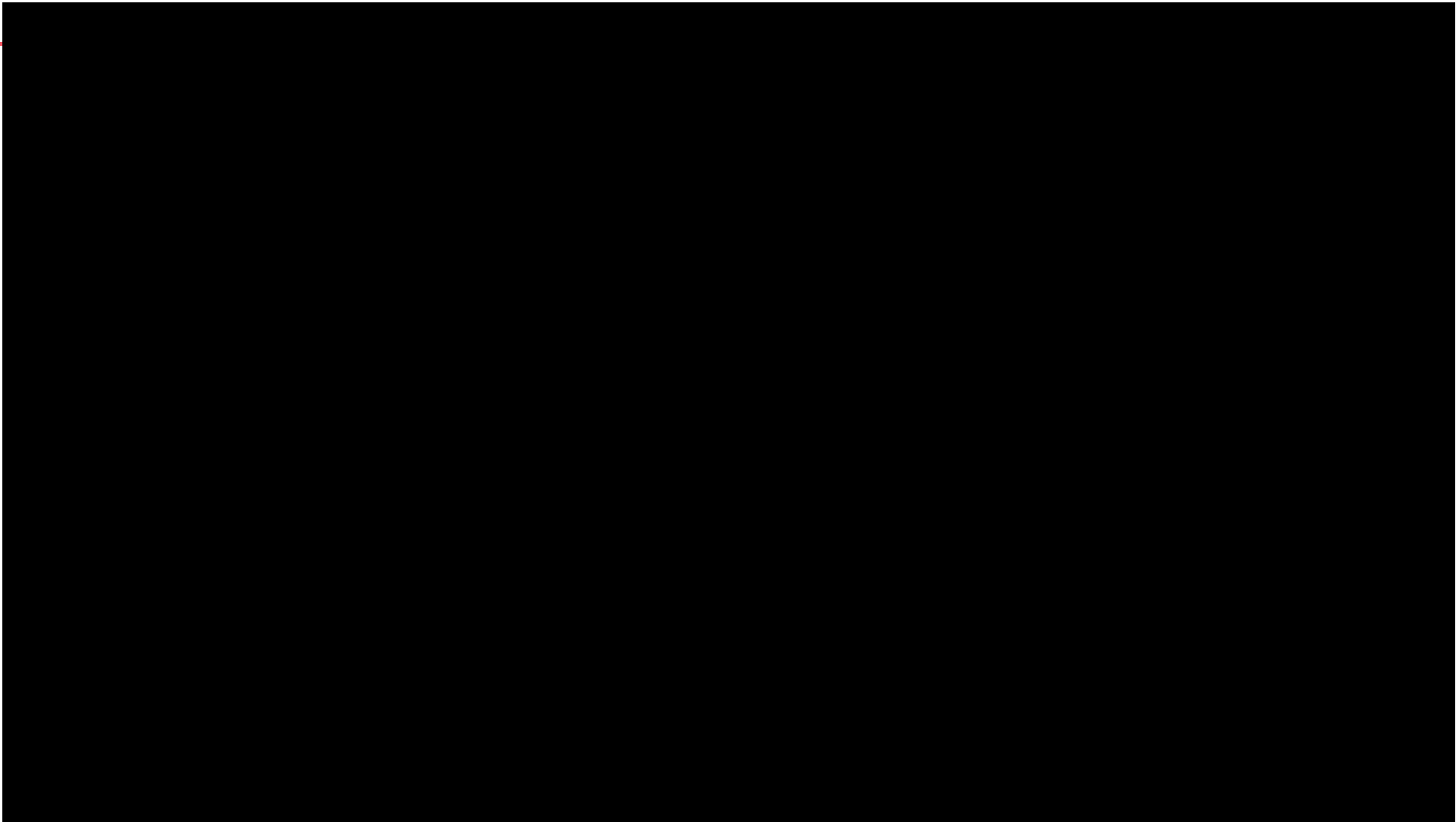
Tagging in Madoc




Adding IIIF content to Madoc

Import and create a shallow copy of
a manifest

Add the manifest to a project



[tei-demo](#)
[Back to site](#) Dashboard Manifests[Import manifest](#)[View manifests with OCR](#) Collections Projects Media Localisation Site configuration Global[Back to site](#) / [Site admin](#) / [Manifests](#)

Manifests

[Import manifest](#)

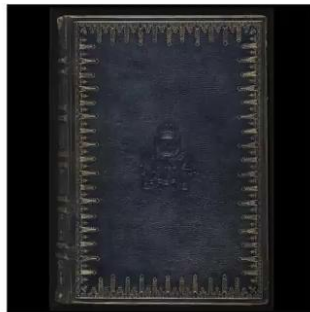
Search manifests

[Search](#)

Page 1 of 1



Lausanne, Bibliothèque canton...
1081 images



[Biblia Latina] (3069 Vol. 1)
681 images



Genesis - Psalmi
671 images



Tagging and Reviewing in Madoc

Tagging

- Select a manifest and canvas
- Autocomplete against the TEI endpoint provided by the Discovery service to select:
 - Book
 - Chapter
 - Verse
- Optionally select a region on the canvas
- Projects can be configured to require a review by a reviewer or administrator
- Typically, a manifest would only get published to the IIIF Change Discovery feed when:
 - All canvases were annotated (so the manifest is complete)
 - All annotations were approved by a reviewer
- For this Demo, I manually published the manifests to the feed early

Projects > TEI

[View in Admin](#)[Project](#)

TEI

Link TEI with IIIF

[Start contributing](#)[Search this project](#)[Reviews](#)[Go to random Manifest](#)[Go to random Canvas](#)

Continue where you left off

Lausanne, Bibliothèque cantonale
et universitaire - Lausanne, U 964

1r

[Continue contribution](#)Lausanne, Bibliothèque cantonale
et universitaire - Lausanne, U 964

1r

[Contribute to the next image](#)[View all contributions](#)

4365

Not started

1

In progress

1

In review

14

Completed

Bodleian Library Arch. B b.10

Photo: © Bodleian Libraries, University of Oxford.
<https://digital.bodleian.ox.ac.uk/terms/>



1



No document yet

Submit for review



Contribution submitted

Keep working on this image or move on to next image



Next image

Back to project

Close and keep working

View in Admin

Project

Manifest

Canvas

Browse all



Page 7 of 668

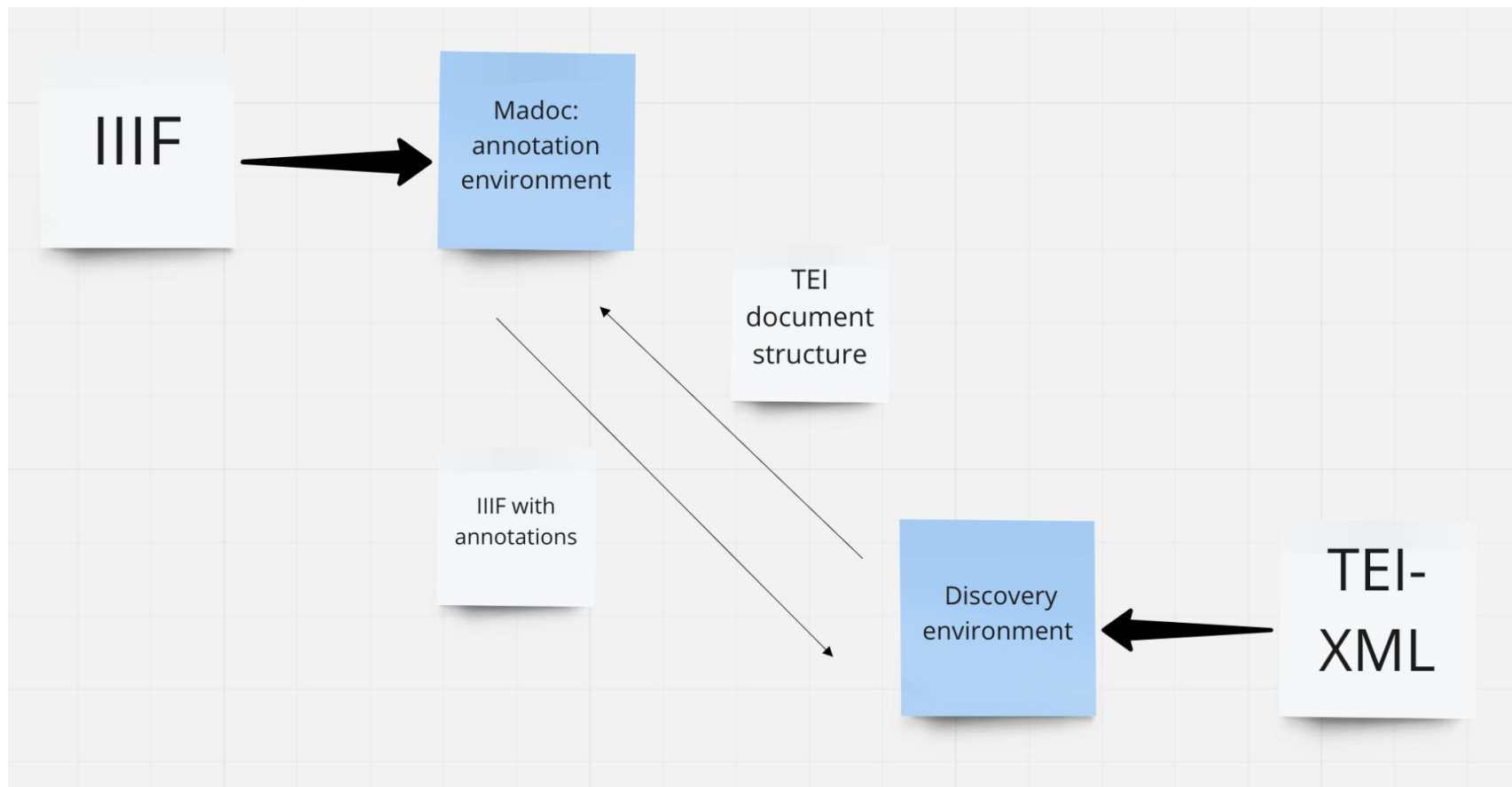


Define region

ment-section

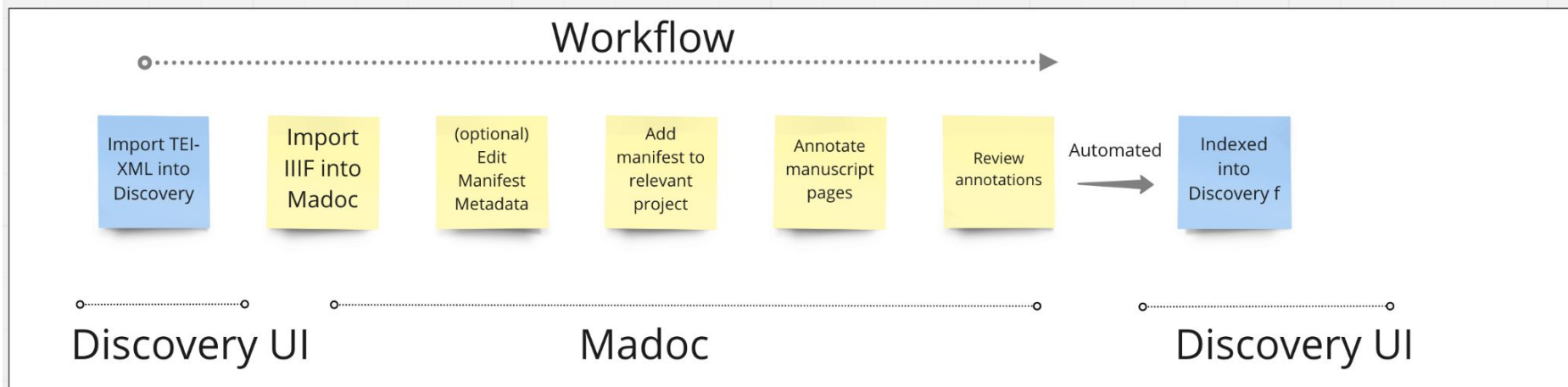
Linking Madoc and Discovery





How does content sync across?

- Madoc publishes a IIIF Change Discovery feed for the content that has been annotated in Madoc
- Discovery UI polls this (see previous slides)
- Add / Update / Delete events lead to the creation or update of IIIF resources in the Discovery service
- Annotations in Madoc are transformed by the Discovery service into Structures/Ranges on IIIF resources (for interoperability)
- Range identifiers are indexed into Search
- Currently, the service creates/stores annotations for the TEI-XML textual content, however ...
- Text display in the Discovery UI is via a custom widget rather than direct from vanilla IIIF compatible annotations



Demo

Challenges, and lessons learned



What worked and was practical for a prototype?

- IIIF provided specifications that were a good fit for:
 - Data modelling the text and image together
 - Modelling document structure in a IIIF environment
 - Synchronising data between two separate systems
- Madoc was a good fit for:
 - Annotation creation environment
 - Review and approval workflow
 - Basic IIIF resource management and metadata editing

Challenges

- Tension between building:
 - A fully generalisable platform
 - A platform that serves the specific data and content for the project concerned
- Site tries to be both
 - UX for navigation is quite tailored to to the requirements / data model for the original manuscript structure
 - However, the underlying data APIs are much more flexible and extensible for future projects / data

Challenges

- Flexibility of TEI-XML as a standard meant that we have to tailor our parsing and indexing of this content around the project's TEI-XML
- Site is quite opinionated about how identifiers should be created in the TEI
- Site tries to be quite flexible about how it handles transcript text, however, we haven't tested it with a huge range of samples other than the specific project TEI-XML transcriptions

Conclusions and potential next steps



Conclusions

Conclusions

- IIIF specifications are a very good fit for this kind of project
 - Not just IIIF Presentation and Image APIs but also
 - IIIF Change Discovery
 - IIIF Content State
- It is possible to bring IIIF together with data from the formats that scholars use, such as TEI-XML
- Annotation or crowdsourcing environments can provide a very powerful “glue” between data sources without laborious cross-conversion of data or re-creation of transcripts, translations, document structure, etc if you can leverage shared identifiers across the IIIF and non-IIIF resources

Potential next steps

Next Steps

- Performance, especially around serialisation and deserialisation of data for rapid transfer between environments
- Generalisability for a wider range of TEI-XML schemas / models
- More flexible UI/UX for data navigation in the Discovery environment
- More flexible administrative/configuration interfaces

Thank you!

