The datahub

Vandermaesen, Matthias
2018

The datahub

De/blending museum data

- Setting the stage
  - In which I’ll describe where we came from
- The Datahub Project
  - In which I’ll show you an aggregation architecture
- The story thus far
  - In which I’ll discuss the construction process
- What we learned
  - In which I’ll conclude with a few take aways

flemishartcollection
MUSEUMS OF FINE ARTS ANTWERP BRUGES GHENT LEUVEN OSTEND

MSK

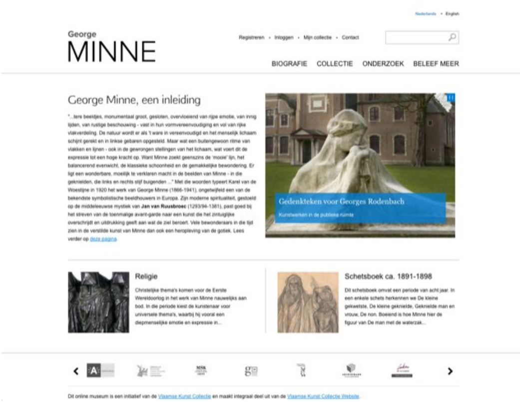
Mu.
ZEE

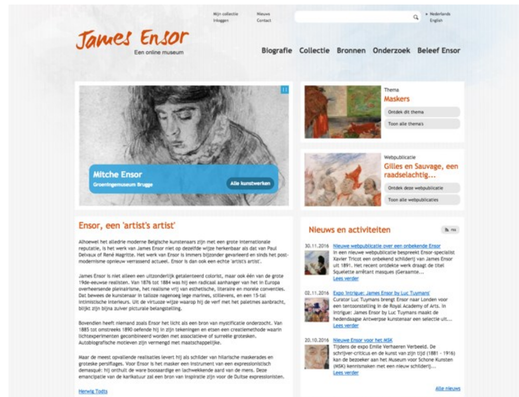
M VAN
MUSEUM
LEUVEN

KONINKLIJK
MUSEUM
VOOR SCHONE
KUNSTEN
ANTWERPEN

BRU
GGE

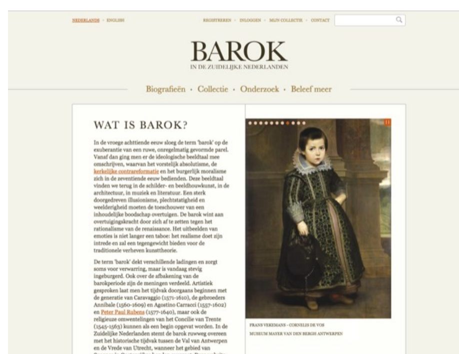
MUSEA
BRUGGE

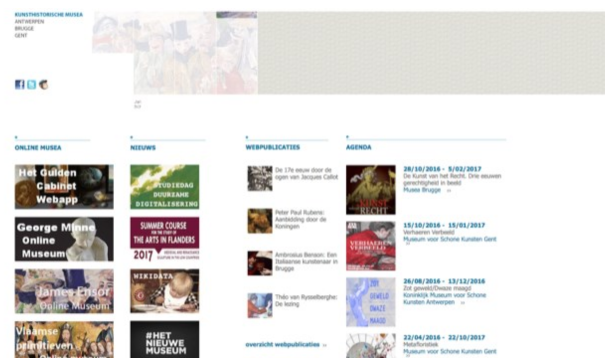


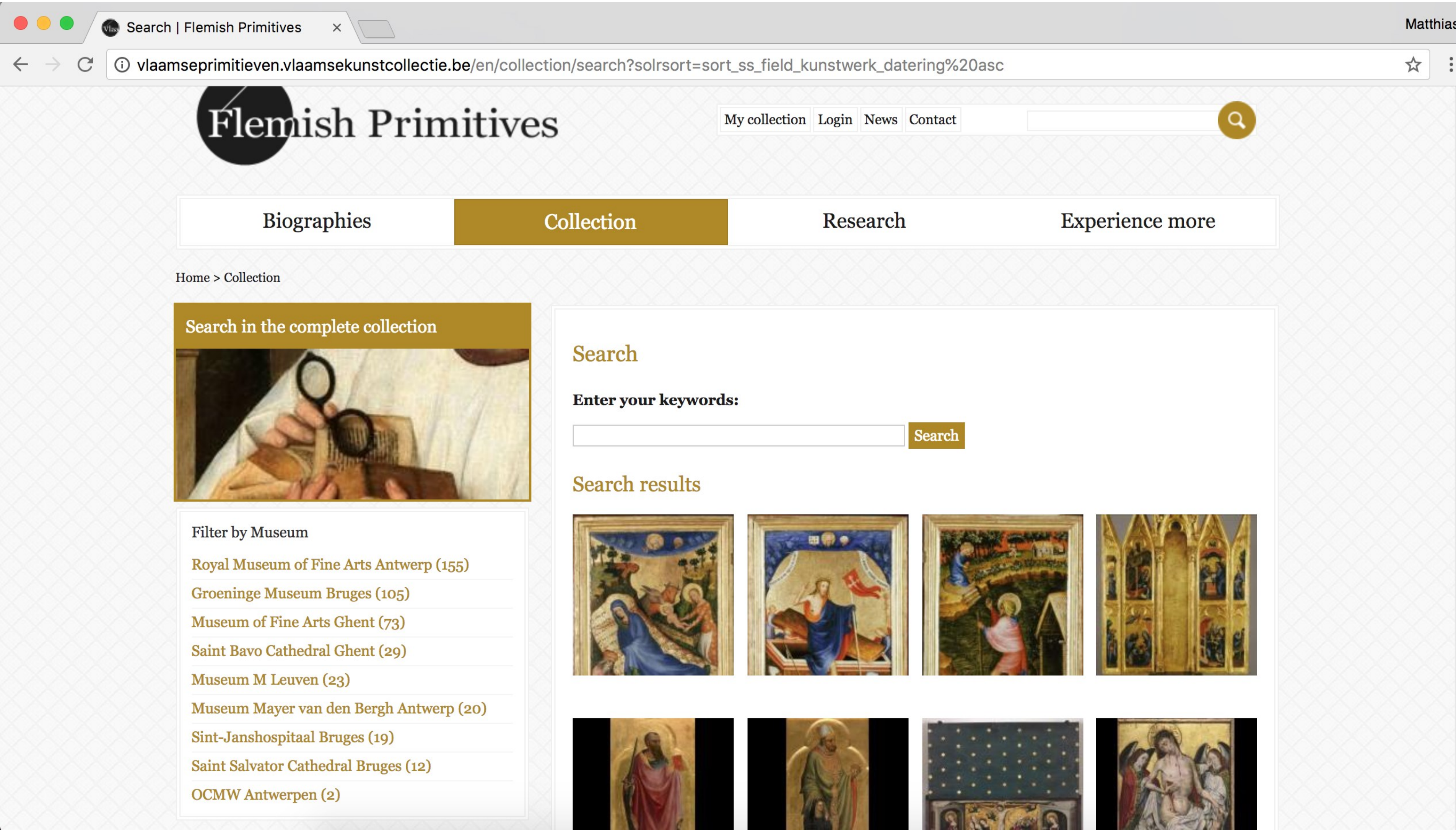


europeana

WIKIDATA






Search | Flemish Primitives
Matthias
vlaamseprimitieven.vlaamsekunstcollectie.be/en/collection/search?solrsort=sort_ss_field_kunstwerk_datering%20asc
Flemish Primitives
My collection
Login
News
Contact
Biographies
Collection
Research
Experience more
Home > Collection
Search in the complete collection
Search
Enter your keywords:
Search
Search results
Filter by Museum
Royal Museum of Fine Arts Antwerp (155)
Groeninge Museum Bruges (105)
Museum of Fine Arts Ghent (73)
Saint Bavo Cathedral Ghent (29)
Museum M Leuven (23)
Museum Mayer van den Bergh Antwerp (20)
Sint-Janshospitaal Bruges (19)
Saint Salvator Cathedral Bruges (12)
OCMW Antwerpen (2)

Multiple organisations

Different local traditions, thesauri, various cataloguing rules (SPECTRUM), organisational contexts,…

Multiple registration systems

TMS, Adlib, CollectiveAccess, closed/open source, Lack of API's, non standardised API's,…

Multiple end user applications

various websites, historically grown, different contractors, various CMS systems, different ways to deliver data,…

Manual exchange

Different ways

- Excel, CSV, vendor formats. WeTransfer, e-mail,..

Error prone

- Corrupt exports, wrong data exported, wrong version passed on, stuff gets lost along the way,…

High overhead costs

- Time and money (communication, $/hour)

High latency

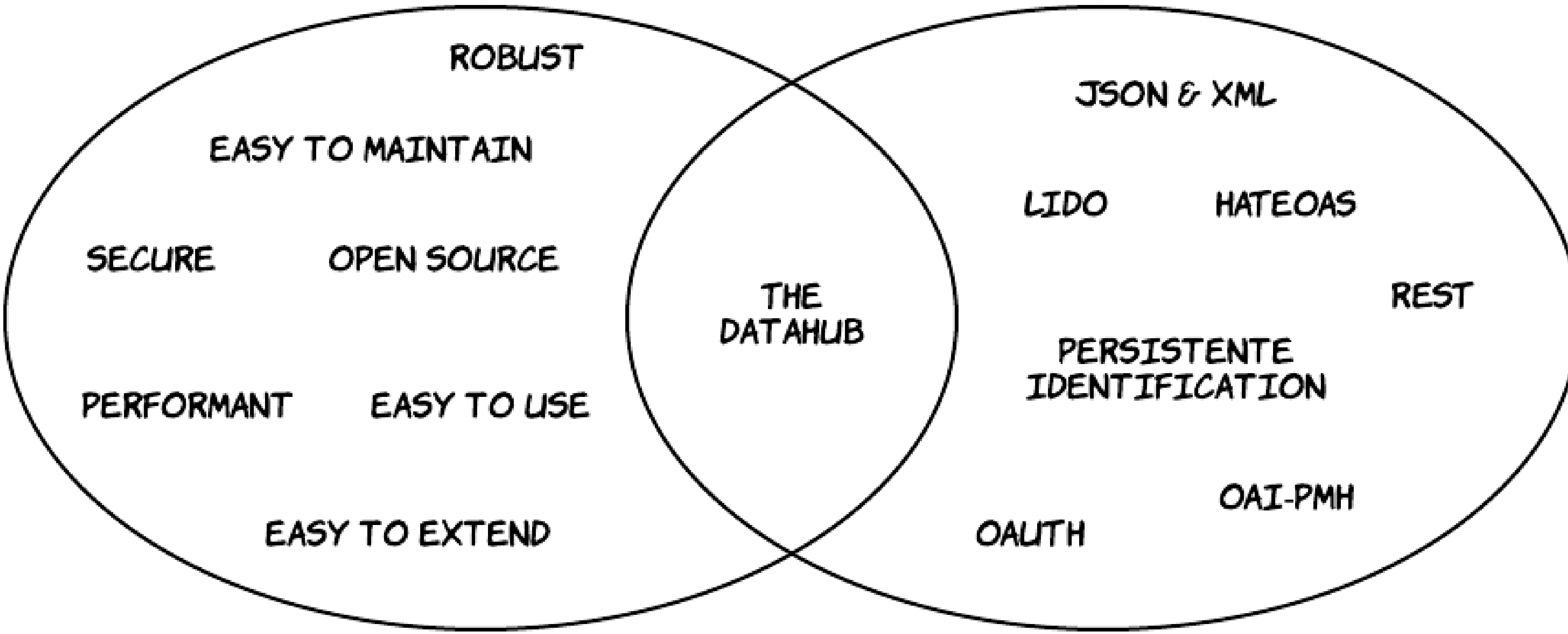
- What’s online is not really up to date

Herding cats

A modern ecosystem

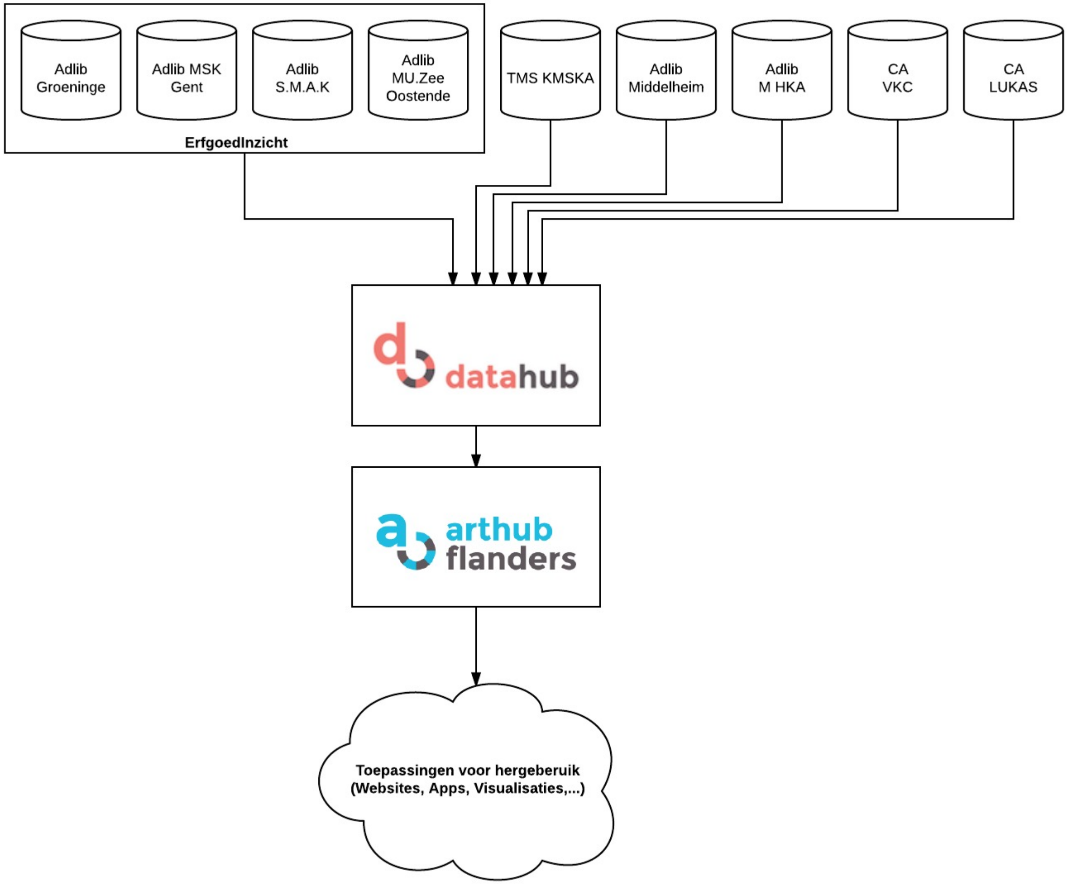
The Datahub Project

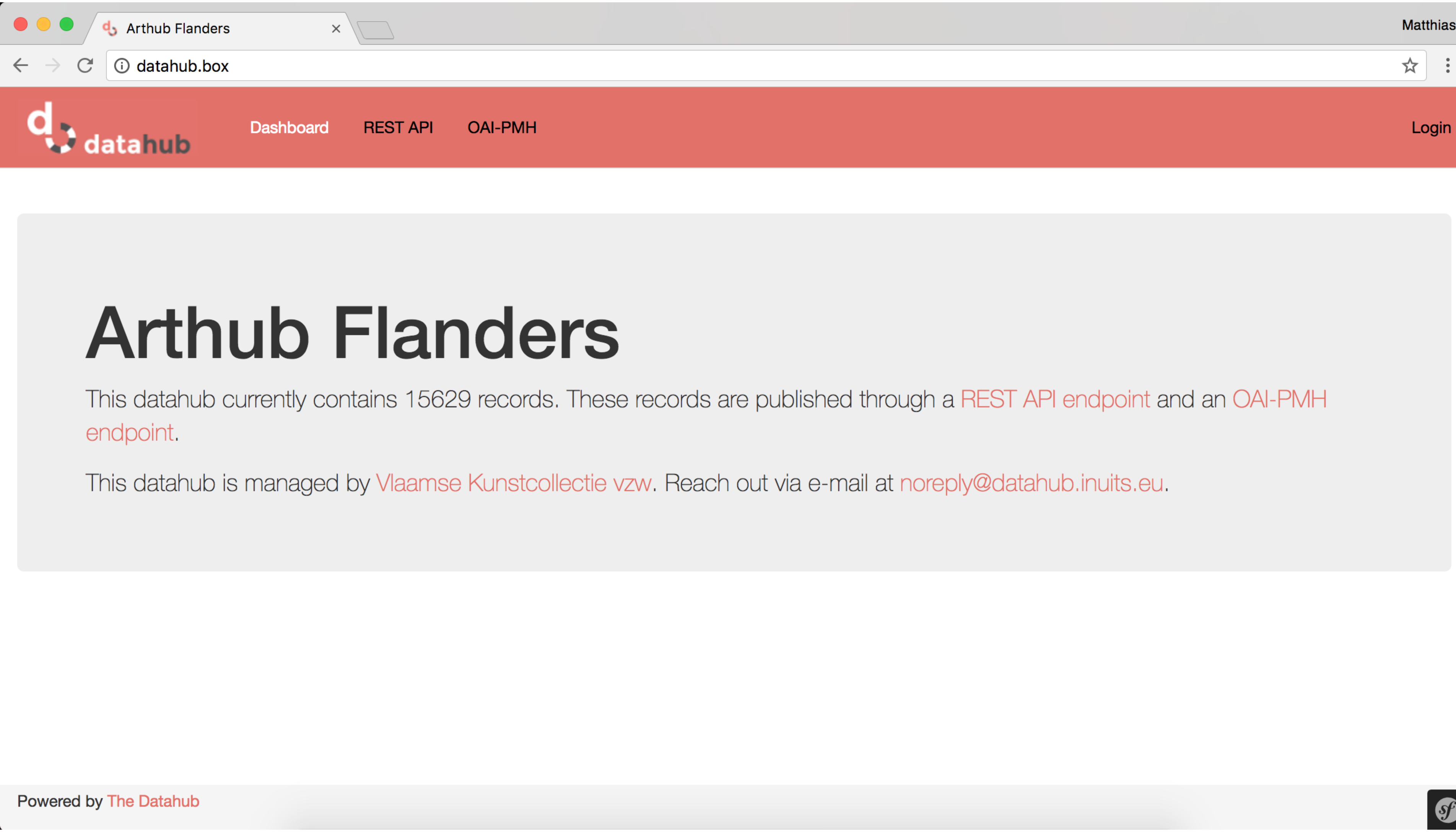
Aggregator



ROBUST
EASY TO MAINTAIN
SECURE
OPEN SOURCE
PERFORMANT
EASY TO USE
EASY TO EXTEND
THE DATAHUB
JSON & XML
LIDO
HATEOAS
REST
PERSISTENTE IDENTIFICATION
OAI-PMH
OAUTH

Local aggregator




Arthub Flanders
Matthias
datahub.box
datahub
Dashboard
REST API
OAI-PMH
Login
Arthub Flanders
This datahub currently contains 15629 records. These records are published through a REST API endpoint and an OAI-PMH endpoint.
This datahub is managed by Vlaamse Kunstcollectie vzw. Reach out via e-mail at noreply@datahub.inuits.eu.
Powered by The Datahub

datahub.box/oai/?metadataPre ×

Matthias

datahub.box/oai/?metadataPrefix=oai_lido&verb=ListRecords

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


```
▼<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2018-06-01T13:15:31Z</responseDate>
  <request metadataPrefix="oai_lido" verb="ListRecords">http://datahub.box</request>
 ▼<ListRecords>
  ▼<record>
   ▼<header>
    ▼<identifier>
       oai:datahub.vlaamsekunstcollectie.be:groeningemuseum.be:0000_GRO1561_I
     </identifier>
     <datestamp>2018-05-02T14:42:04Z</datestamp>
   </header>
   ▼<metadata>
    ▼<lido:lido xmlns:gml="http://www.opengis.net/gml" xmlns:lido="http://www.lido-schema.org" xmlns:xlink="http://www.w3.org/1999/xlink">
     ▼<lido:lidoRecID lido:pref="alternate" lido:type="purl" lido:source="Musea Brugge - Groeningemuseum" lido:label="dataPID">
        http://groeningemuseum.be/collection/work/data/0000_GRO1561_I
      </lido:lidoRecID>
     ▼<lido:lidoRecID lido:pref="preferred" lido:type="urn" lido:source="Vlaamse Kunstcollectie - Arthub Flanders" lido:label="dataPID">
        oai:datahub.vlaamsekunstcollectie.be:groeningemuseum.be:0000_GRO1561_I
      </lido:lidoRecID>
     ▼<lido:objectPublishedID lido:type="purl" lido:source="Musea Brugge - Groeningemuseum" lido:label="workPID">
        http://groeningemuseum.be/collection/work/id/0000_GRO1561_I
      </lido:objectPublishedID>
     ▼<lido:category>
        <lido:conceptID lido:type="purl" lido:source="cidoc-crm">http://www.cidoc-crm.org/crm-concepts/E22</lido:conceptID>
        <lido:term>Man-Made Object</lido:term>
      </lido:category>
     ▼<lido:descriptiveMetadata xml:lang="nl">
      ▼<lido:objectClassificationWrap>
       ▼<lido:objectWorkTypeWrap>
        ▼<lido:objectWorkType>
           <lido:conceptID lido:pref="preferred" lido:type="local" lido:source="Adlib">20000001</lido:conceptID>
           <lido:conceptID lido:pref="alternate" lido:type="purl" lido:source="AAT">http://vocab.getty.edu/aat/300033618</lido:conceptID>
           <lido:term lido:pref="preferred" xml:lang="nl">schilderingen</lido:term>
           <lido:term lido:pref="alternate" xml:lang="nl">schilderingen</lido:term>
         </lido:objectWorkType>
        </lido:objectWorkTypeWrap>
       ▼<lido:classificationWrap>
        ▼<lido:classification>
           <lido:conceptID lido:pref="preferred" lido:type="local" lido:source="Adlib">20000153</lido:conceptID>
```

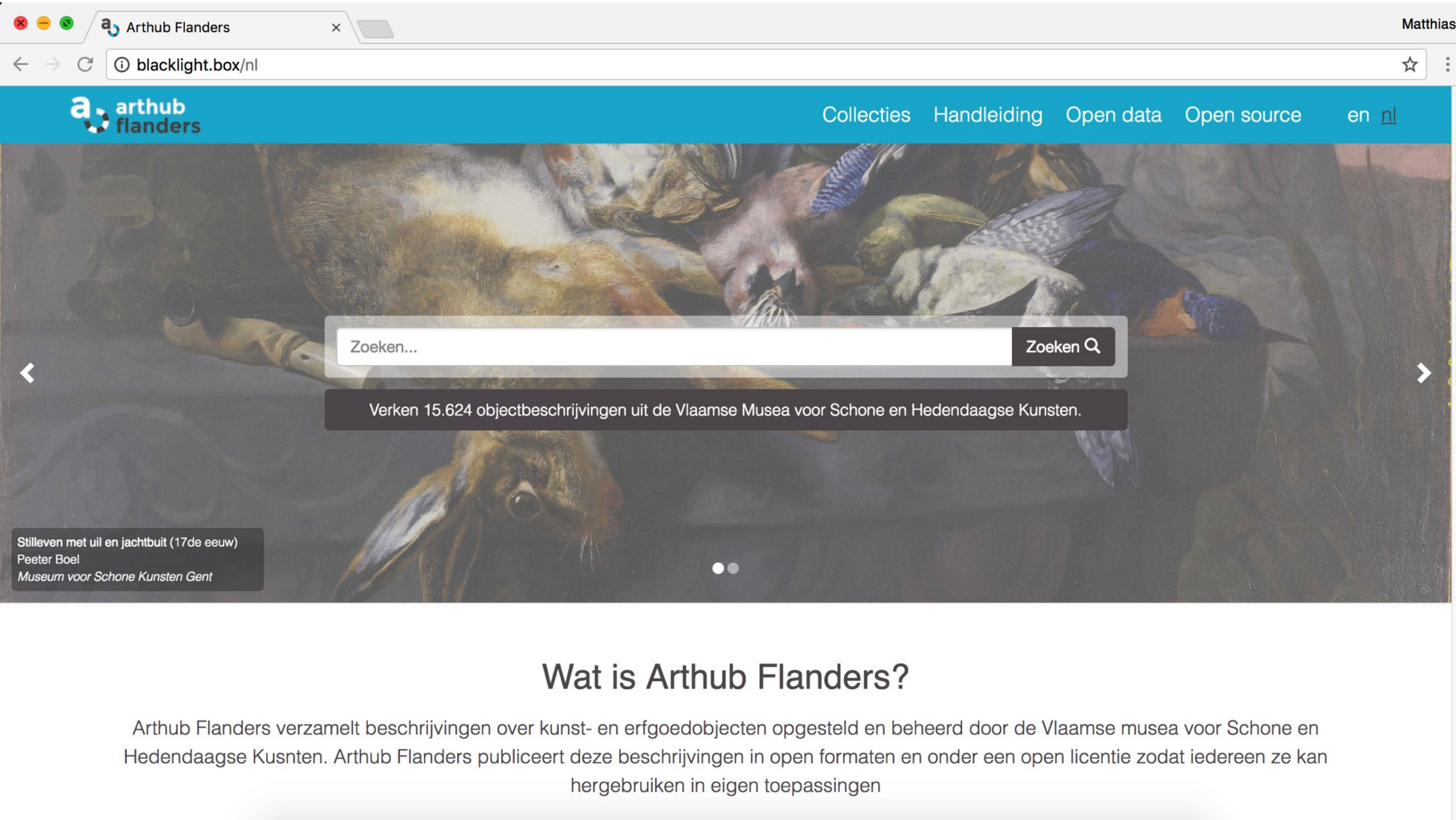
datahub.box/api/v1/data.json

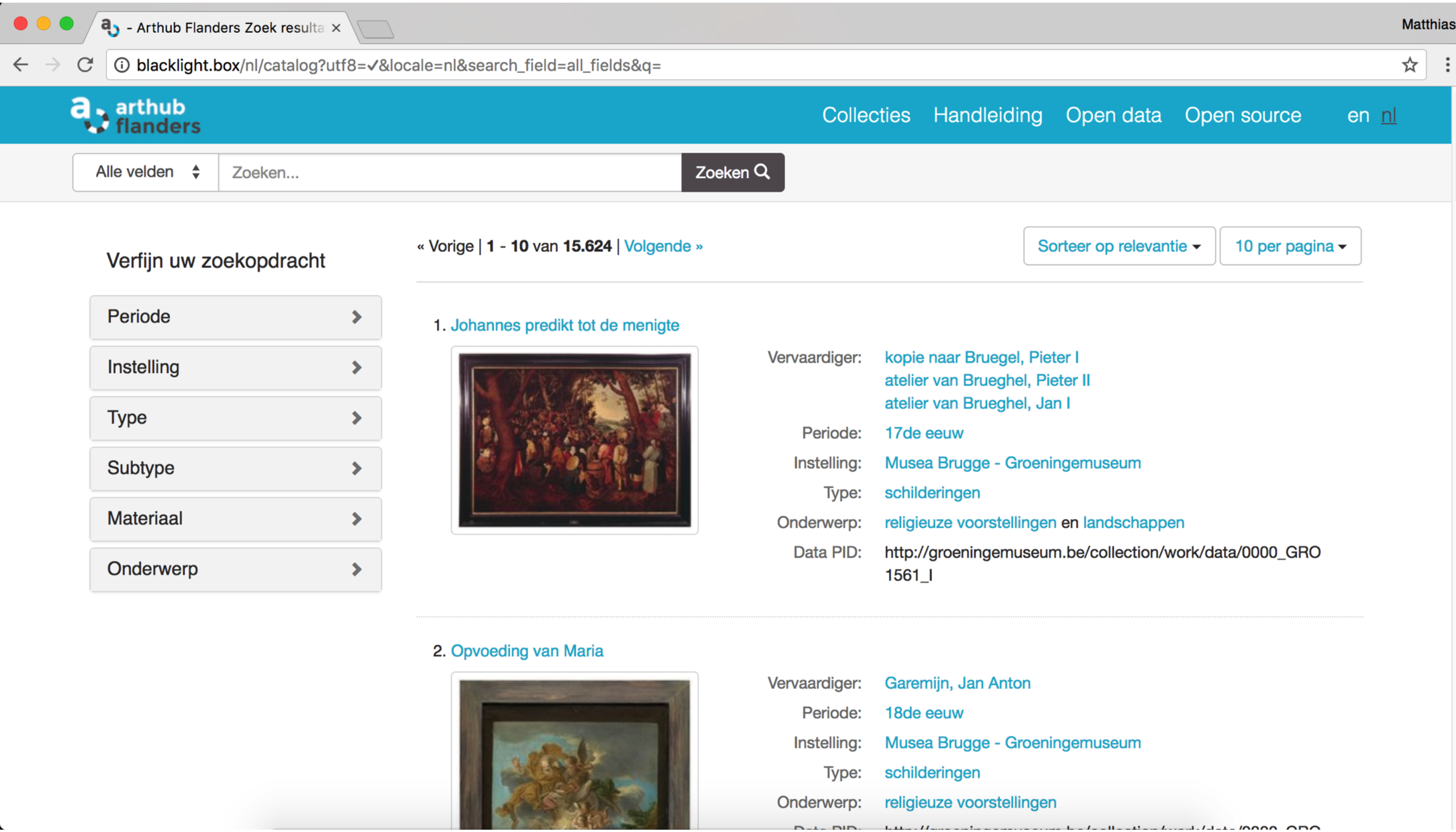
Matthias

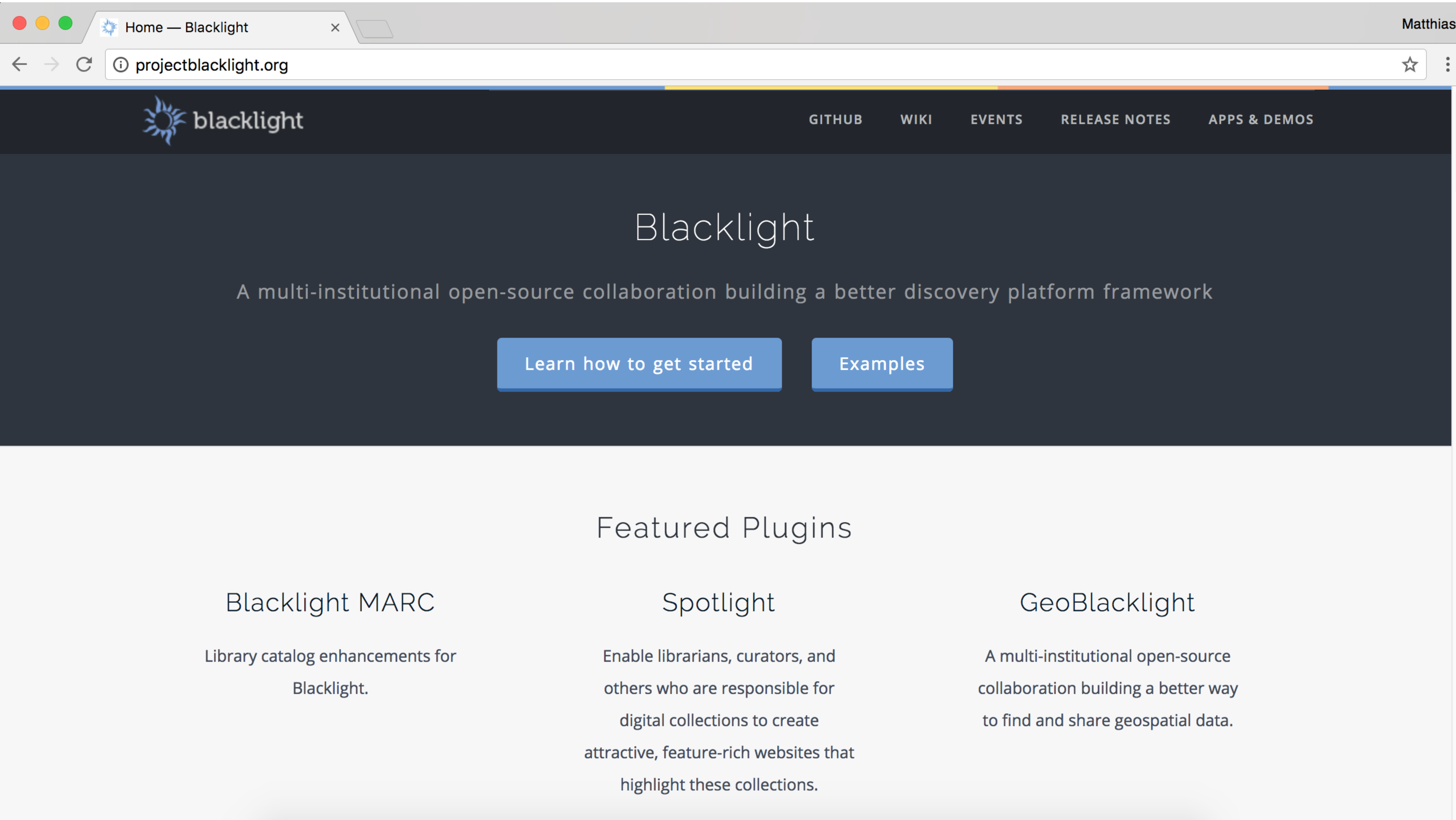
datahub.box/api/v1/data.json

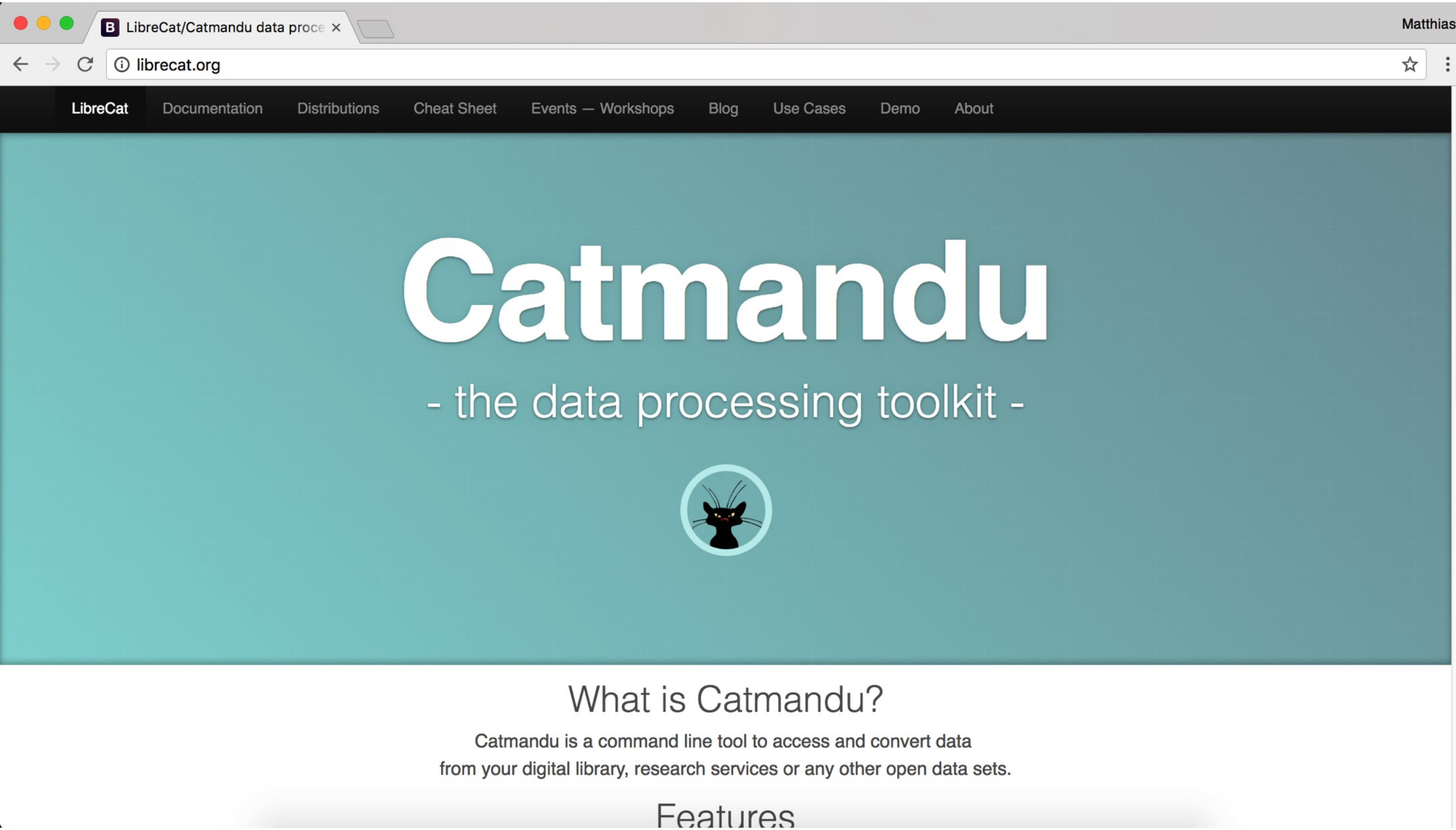

```
// 20180601151621
// http://datahub.box/api/v1/data.json

{
  "offset": 0,
  "limit": 5,
  "total": 15629,
  "_links": {
    "self": {
      "href": "/api/v1/data?limit=5"
    },
    "first": {
      "href": "/api/v1/data?limit=5"
    },
    "last": {
      "href": "/api/v1/data?offset=15625&limit=5"
    },
    "next": {
      "href": "/api/v1/data?offset=5&limit=5"
    }
  },
  "_embedded": {
    "records": Array[5][
      {
        "id": "5ae9ce3c72a84303de6f2ada",
        "created": "2018-05-02T09:42:04-05:00",
        "updated": "2018-05-02T09:42:04-05:00",
        "json": Array[6][
          {
            "name": "{http://www.lido-schema.org}lidoRecID",
            "value": "http://groeningemuseum.be/collection/work/data/0000_GRO1561_I",
            "attributes": {
              "{http://www.lido-schema.org}pref": "alternate",
              "{http://www.lido-schema.org}type": "purl",
              "{http://www.lido-schema.org}source": "Musea Brugge - Groeningemuseum",
              "{http://www.lido-schema.org}label": "dataPID"
```


Arthub Flanders
Matthias
blacklight.box/nl
arthub flanders
Collecties
Handleiding
Open data
Open source
en nl
Zoeken...
Zoeken
Verken 15.624 objectbeschrijvingen uit de Vlaamse Musea voor Schone en Hedendaagse Kunsten.
Stilleven met uil en jachtbuit (17de eeuw)
Peeter Boel
Museum voor Schone Kunsten Gent
Wat is Arthub Flanders?
Arthub Flanders verzamelt beschrijvingen over kunst- en erfgoedobjecten opgesteld en beheerd door de Vlaamse musea voor Schone en Hedendaagse Kusnten. Arthub Flanders publiceert deze beschrijvingen in open formaten en onder een open licentie zodat iedereen ze kan hergebruiken in eigen toepassingen


- Arthub Flanders Zoek resulta
Matthias
blacklight.box/nl/catalog?utf8=✓&locale=nl&search_field=all_fields&q=
arthub flanders
Collecties
Handleiding
Open data
Open source
en nl
Alle velden
Zoeken...
Zoeken
Verfijn uw zoekopdracht
Periode
Instelling
Type
Subtype
Materiaal
Onderwerp
« Vorige | 1 - 10 van 15.624 | Volgende »
Sorteer op relevantie
10 per pagina
1. Johannes predikt tot de menigte
Vervaardiger: kopie naar Bruegel, Pieter I
atelier van Brueghel, Pieter II
atelier van Brueghel, Jan I
Periode: 17de eeuw
Instelling: Musea Brugge - Groeningemuseum
Type: schilderingen
Onderwerp: religieuze voorstellingen en landschappen
Data PID: http://groeningemuseum.be/collection/work/data/0000_GRO1561_I
2. Opvoeding van Maria
Vervaardiger: Garemijn, Jan Anton
Periode: 18de eeuw
Instelling: Musea Brugge - Groeningemuseum
Type: schilderingen
Onderwerp: religieuze voorstellingen


Home — Blacklight
Matthias
projectblacklight.org
blacklight
GITHUB
WIKI
EVENTS
RELEASE NOTES
APPS & DEMOS
Blacklight
A multi-institutional open-source collaboration building a better discovery platform framework
Learn how to get started
Examples
Featured Plugins
Blacklight MARC
Library catalog enhancements for Blacklight.
Spotlight
Enable librarians, curators, and others who are responsible for digital collections to create attractive, feature-rich websites that highlight these collections.
GeoBlacklight
A multi-institutional open-source collaboration building a better way to find and share geospatial data.



LibreCat/Catmandu data proce
Matthias
librecat.org
LibreCat
Documentation
Distributions
Cheat Sheet
Events — Workshops
Blog
Use Cases
Demo
About
Catmandu
- the data processing toolkit -
What is Catmandu?
Catmandu is a command line tool to access and convert data
from your digital library, research services or any other open data sets.
Features

demo.ini UNREGISTERED

demo.ini

```ini
[General]
id_path = '_metadata.object_number'

[Importer]
plugin = EIZ

[plugin_importer_EIZ]
endpoint = https://endpoint.org
metadata_prefix = oai_adlib
username = username
password = secretpassword
handler = adlib

[Fixer]
plugin = Fix

[plugin_fixer_Fix]
condition_path = "_metadata.institution\\.name.value"
fixers = MSK, GRO

[plugin_fixer_MSK]
condition = 'Museum voor Schone Kunsten Gent'
file_name = '/Users/matthiasvandermaesen/Workspace/Datahub-Fixes/msk_oai_adlib.fix'

[plugin_fixer_GRO]
condition = 'Musea Brugge - Groeningemuseum'
file_name = '/Users/matthiasvandermaesen/Workspace/Datahub-Fixes/groeninge_oai_adlib.fix'

[Exporter]
plugin = Arthub

[plugin_exporter_Arthub]
datahub_url = http://datahub.box
datahub_format = LIDO
oauth_client_id = slightlylesssecretpublicid
oauth_client_secret = supersecretsecretphrase
oauth_username = admin
oauth_password = datahub
```

Line 1, Column 1 Tab Size: 4 Plain Text

groeninge_oai_adlib.fix UNREGISTERED

groeninge_oai_adlib.fix

```
### LIDO lidoRecID

    # ID
    #
    # The ID in Solr is based on the data_pid. The data_pid is converted to a string
    # which can be safely used as an identifier in Project Blacklight. The format of
    # the ID field looks like this:
    #
    #   oai:datahub.vlaamsekunstcollectie.be:<domain>:<identifier>
    #   ex. oai:datahub.vlaamsekunstcollectie.be:kmksa.be:254
    #   ex. oai:datahub.vlaamsekunstcollectie.be:collectievlaamsegemeenschap.be:837
    #
    # Note: the .tld is stripped from the domainname because the . (dot) breaks the
    # route matching algoritm.

    unless is_array(or_record.dataPid)
        move_field(or_record.dataPid, or_record.tmp)
        set_array(or_record.dataPid)
        move_field(or_record.tmp, or_record.dataPid.$last)
    end

    do list(path:or_record.dataPid, var: c)

        if all_match('c.digital_reference\.description.value', 'datapid')

            lido_baseid(
                lidoRecID,
                c.digital_reference,
                -type: purl,
                -source: 'Musea Brugge - Groeningemuseum',
                -label: dataPID,
                -pref: alternate
            )

            copy_field(c.digital_reference, or_record.oaiPid)
            parse_text(or_record.oaiPid, '.*://([A-Za-z0-9\-\.]+)/collection/work/data/(.*)')
            join_field('or_record.oaiPid', ':')
            prepend('or_record.oaiPid', 'oai:datahub.vlaamsekunstcollectie.be:')

            lido_baseid(
                lidoRecID,
                or_record.oaiPid,
```

Line 1, Column 1 Spaces: 4 Plain Text

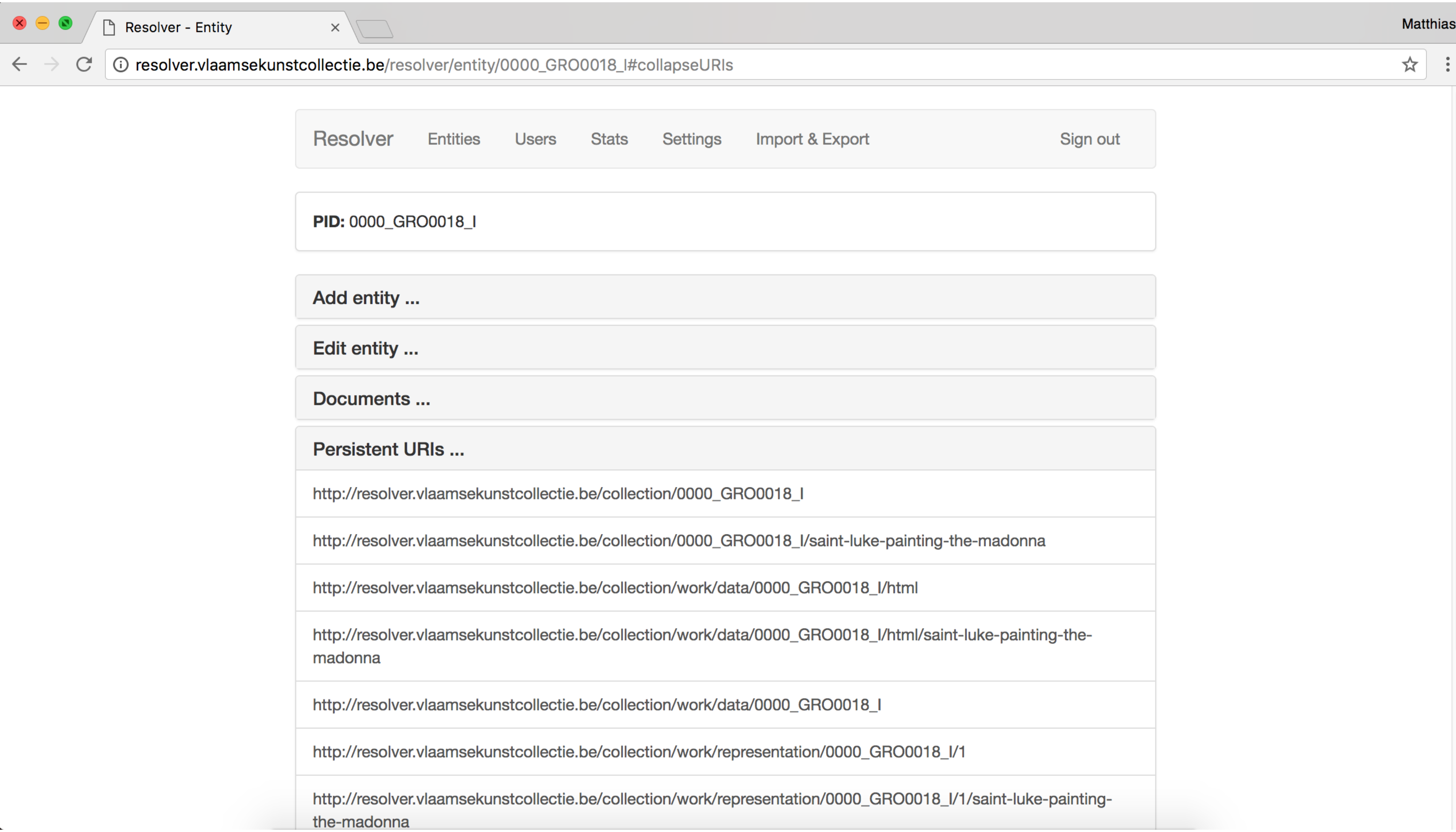
2. ~ (bash)

```
matthiasvandermaesen at Artemis in ~
$ dhconveyor
   commands: list the application's commands
       help: display a command's help screen

      index: Transport data from a flat file to a data index in bulk.
  transport: Transport data from a data source to a data sink.


matthiasvandermaesen at Artemis in ~
$ dhconveyor transport -p ~/Workspace/arthub-pipelines/erfgoedinzicht.ini -v
Loading pipeline configuration...
Initializing importer/exporter...
Initializing fixers...
Importing data from source...
✔ - Item #1 : 0000.GR01561.I (id): exported.
✔ - Item #2 : 0000.GR01390.I (id): exported.
✔ - Item #3 : 0000.GR00128.I (id): exported.
✔ - Item #4 : 0000.GR01476.I (id): exported.
✔ - Item #5 : 0000.GR00479.I (id): exported.
✔ - Item #6 : 0000.GR01372.I (id): exported.
✔ - Item #7 : 0000.GR01360.I (id): exported.
✔ - Item #8 : 0000.GR01359.I (id): exported.
✔ - Item #9 : 0000.GR00227.I (id): exported.
✔ - Item #10 : 0000.GR01280.I (id): exported.
✔ - Item #11 : 0000.GR01243.I (id): exported.
✔ - Item #12 : 0000.GR00299.I (id): exported.
✔ - Item #13 : 0000.GR01230.I (id): exported.
^C

matthiasvandermaesen at Artemis in ~
$
```


Resolver - Entity
Matthias
resolver.vlaamsekunstcollectie.be/resolver/entity/0000_GRO0018_I#collapseURIs
Resolver
Entities
Users
Stats
Settings
Import & Export
Sign out
PID: 0000_GRO0018_I
Add entity ...
Edit entity ...
Documents ...
Persistent URIs ...
http://resolver.vlaamsekunstcollectie.be/collection/0000_GRO0018_I
http://resolver.vlaamsekunstcollectie.be/collection/0000_GRO0018_I/saint-luke-painting-the-madonna
http://resolver.vlaamsekunstcollectie.be/collection/work/data/0000_GRO0018_I/html
http://resolver.vlaamsekunstcollectie.be/collection/work/data/0000_GRO0018_I/html/saint-luke-painting-the-madonna
http://resolver.vlaamsekunstcollectie.be/collection/work/data/0000_GRO0018_I
http://resolver.vlaamsekunstcollectie.be/collection/work/representation/0000_GRO0018_I/1
http://resolver.vlaamsekunstcollectie.be/collection/work/representation/0000_GRO0018_I/1/saint-luke-painting-the-madonna

The story thus far

Assumptions / Reality

- We had a fixed, limited budget
- Estimated timeline 3 to 6 months.
- A production ready version.

- Contractor delivered a prototype version.
- We over-extended the timing.
- Switch to DIY development after 6 months.
- Scope changes as we went along.

What happened?

- We underestimated the ETL workload
- We overestimated contractor engagement
- We underestimated organisational complexity

Wicked ETL

- Context really matters
- Getting intimate with the domain takes time
- Integrating data across network is challenging.
- Difficult to guestimate complexity up front

Wicked ETL

Context really matters

- Machines
  Legacy software, lack of infrastructure,…
- People
  Data means nothing until it gets interpreted.
  But, different perceptions of reality…
- Content
  Driven by tradition, software, people.

Wicked ETL

Data modelling is a wicked challenge

- Mapping to standardised exchange formats
  … and their specific data models
- Normalisation and enrichment
  … are we taking about the same thing?
- Context specific concerns
  ... Copyright, privacy, security, authority

Procurement

- Build-to-print vs build-to-spec.
- You outsource the process, not the project.
- Is contractor service a good fit?
- Relationship with the contractor!
- Procurement is part of the design process

DIY development

- Knowledge domain and technical experience
- Flexibility to build exactly what you need
- Reduces dependency on a specific contractor
- Requires in-house competences
- Payroll is a hidden cost
- The Bus Factor risk

Lessons learned

- Own your project
  - Define the project process you’re going to follow
- Actively involve your stakeholders
  - Challenge your own assumptions, but keep your focus!
- Actively be involved in the process
  - Don’t assume a vendor will solve things for you.

- Be mindful about the budget
  - Fixed price vs Fixed buget
- In source talented specialists you need
  - Identify right profile: IA, Dev, PM, UX,…
- Outsource placing the kitchen sink
  - Stock off-the-shelf website or web app

- Document all the things
  - Be mindful about the human who comes after you!
- Don’t do elaborate specifications up front
  - Nobody is interested in paper tigers.
- Make your hands dirty
  - Try tools up front. Identify the big hurdles early.

Thank you!

https://github.com/thedatahub

https://thedatahub.github.io

http://www.flemishartcollection.be

T: @netsensei