The datahub
De/blending museum data

Data has a better idea
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
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

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

```xml
  <responseDate>2018-06-01T13:15:31Z</responseDate>
  <request metadataPrefix="oai_lido" verb="ListRecords">http://datahub.box</request>
</OAIPMH>
```
{
    "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_GR01561_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"
                        }
                    }
                ]
            }
        ]
    }
}
Wat is Arthub Flanders?

Arthub Flanders verzamelt beschrijvingen over kunst- en erfgoedobjecten opgesteld en beheerd door de Vlaamse musea voor Schone en Hedendaagse Kunsten. Arthub Flanders publiceert deze beschrijvingen in open formaten en onder een open licentie zodat iedereen ze kan hergebruiken in eigen toepassingen.
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: schilderijen
Onderwerp: religieuze voorstellingen en landschappen
Data PID: http://groeningemuseum.be/collection/work/data/0000_GRO1561_I

2. Opvoeding van Maria

Vervaardiger: Garenijn, Jan Anton

Periode: 18de eeuw
Instelling: Musea Brugge - Groeningemuseum
Type: schilderijen
Onderwerp: religieuze voorstellingen
Data PID: http://groeningemuseum.be/collection/work/data/0000_GRO1561_I
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.
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
[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"
filters = 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
## 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:

```plaintext
# oai:database.vlaamsecollectie.be:<domain>:<identifier>
ex. oai:database.vlaamsecollectie.be:kmksa:254
ex. oai:database.vlaamsecollectie.be:collectievlaamsegemeenschap.be:837
```

Note: the .tld is stripped from the domain name because the . (dot) breaks the route matching algorithm.

```plaintext
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
```

```plaintext
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:database.vlaamsecollectie.be:')

lido_baseid(
  lidoRecID,
  or_record.oaiPid,
)
$ 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.

$ dhconveyor transport -p ~/Workspace/arthur-pipelines/erfgoedinzicht.ini -v
Loading pipeline configuration...
Initializing importer/exporter...
Initializing fixers...
Importing data from source...
✓ Item #1: 0000.GROI561.I (id): exported.
✓ Item #2: 0000.GROI390.I (id): exported.
✓ Item #3: 0000.GROI128.I (id): exported.
✓ Item #4: 0000.GROI1476.I (id): exported.
✓ Item #5: 0000.GROI0479.I (id): exported.
✓ Item #6: 0000.GROI1372.I (id): exported.
✓ Item #7: 0000.GROI1360.I (id): exported.
✓ Item #8: 0000.GROI1359.I (id): exported.
✓ Item #9: 0000.GROI0227.I (id): exported.
✓ Item #10: 0000.GROI1280.I (id): exported.
✓ Item #11: 0000.GROI1243.I (id): exported.
✓ Item #12: 0000.GROI0299.I (id): exported.
✓ Item #13: 0000.GROI1230.I (id): exported.
^C
PID: 0000_GRO0018_I

Add entity ...

Edit entity ...

Documents ...

Persistent URIs ...

http://resolver.vlaamsekunstcollectie.be/collection/0000_GRO0018_I


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
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 budget

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