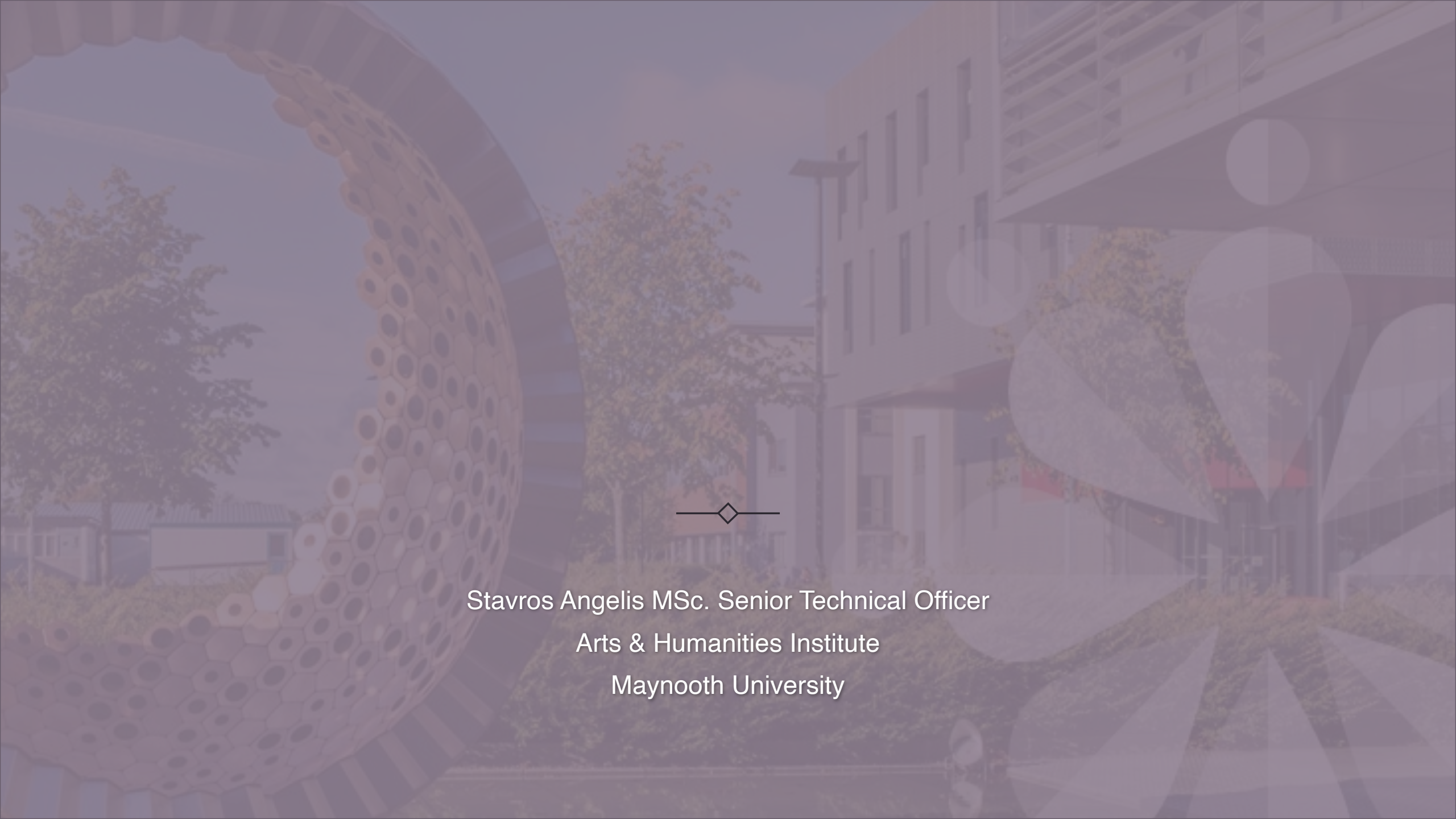


Stavros Angelis MSc. Senior Technical Officer
Arts & Humanities Institute
Maynooth University

Presentation plan

- Introduction to DH projects
- Clericus
  - Research question
  - Work packages
  - Data modelling
  - Live system

A typical DH project

- It all starts with a research question
  - news ways to explore / query the data
  - new ways to present / visualise / interact with the data
  - extending an existing collection
  - …
- And one or more data collections
  - physical collection, library / archival / private etc.
    - unstructured data / semi-structured data
  - digital collection, mostly 2.0 versions of existing projects
    - semi-structured data / structured data
    - need to be migrated to more up-to-date formats

Roles

- Domain experts
  - PIs, researchers, research assistants, postDocs, PhDs …
- Technical personnel
  - programmers, data analysts …
- Other roles
  - project management, communication and outreach, commercialisation …

The project

The project lifecycle is not always the same

- collect data
  - locate and identify data in archives, libraries, digital collections …
  - create a data gathering plan
- prepare data
  - normalise, clean, deduplicate…
- develop a system and ingest data into a database
  - create an ingestion pipeline
- manipulate data in the system
  - update, enrich manually, semi-automatically, automatically
- outputs
  - reports, publications, digital collections, websites, tools, services, products…

Goals

The main goal is to advance research in a certain domain

- Answer existing research questions in new ways
- Be able to ask new research questions
- Visualise data in new ways to gain new insights
- Identify / add extra dimensions to the data (e.g. time and space)
- Make collections public
- Create tools and services for the community
- …

Maynooth University National University of Ireland Maynooth | Arts and Humanities Institute

Clericus

Clericus is a digital humanities research project that aims to develop a database on the Irish clerical population for the early modern and modern periods. Phase I of the project, financed by St. Patrick's College Maynooth (SPCM), focused on the students and faculty of St. Patrick's College, historically Ireland's largest seminary and pontifical university. The principal sources used were class portraits and student lists compiled by Monsignor Patrick J. Hamell and Cora Fennelly. This combined dataset provides just under 20,000 individual biographical entries. In line with data protection recommendations, all class portraits and biographical entries up to and including 1945 are available via the website's database (48 class portraits and approximately 14,000 entries).

Maynooth University | National University of Ireland Maynooth | Arts and Humanities Institute

Research questions and data collections

- The initial research goal was to digitise the classpieces from the St. Patrick's College
- extract information related to the people depicted on the classpieces, organise it, put into a database and publish
- include other data sources and combine / relate the data
- present the results in a publicly available website
- add custom visualisations and explore additional data relationships

Work Packages

- Content WP
  - Collections survey
  - Artefact identification & collection
  - Data extraction & ingestion
  - Data enrichment
- Technical WP
  - Information workflow
  - Data modelling
  - Technical infrastructure development

Maynooth University | Arts and Humanities Institute

Primary source - Classpiece

- A class portrait depicting a year's students, the professor and guests of honor
- Complex item that contains multiple entities and relations

Content Work Package

- Collections survey
- Artefact identification & collection (digitisation)
- Data extraction & ingestion
- Data enrichment

Maynooth University
National University of Ireland Maynooth
Arts and Humanities Institute

Collection survey

- As part of the content identification we decided to run a paraller task, a collections survey:
  - locate collections of interest to the project
  - identify information about the collections according to Dublin Core Collections Application Profile
  - We stored the collected data into a private system, “Sources registry”

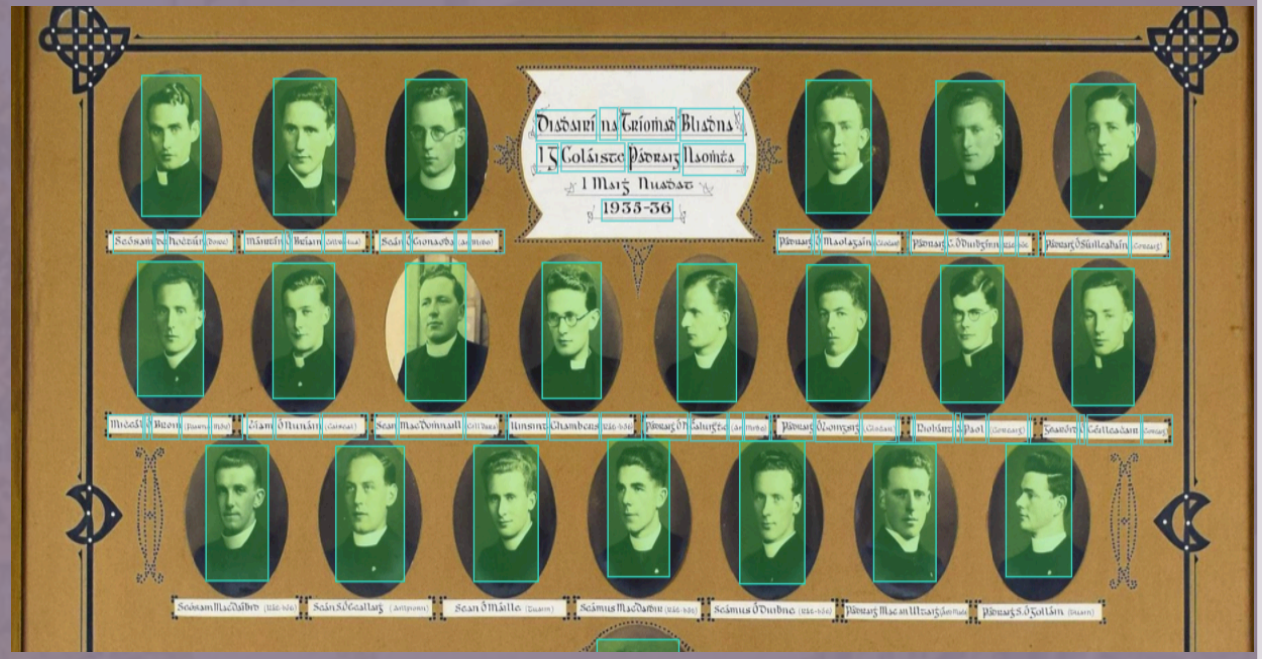
Artefact identification & digitisation

- Locating classpieces on host premises
- Coordinating with host's staff for access / digitisation conditions
- 124 classpieces in total – 1861, 1878, 1880s to 2018
- Photographed in situ or relocated – issues of access/ lighting / equipment
- Captured files in RAW format
- Issues included:
  - uneven walls & floors
  - photograph enhancement (crop, deskew, color / contrast balance etc.)
  - output file of 5MB or less

Maynooth University National University of Ireland Maynooth | Arts and Humanities Institute

Data extraction & ingestion

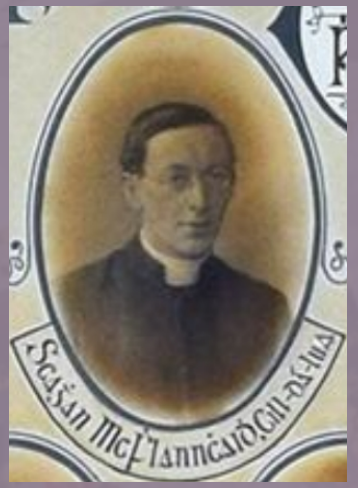
- Image files uploaded to system backend
- Semi-automatic parsing / data extraction with Computer Vision
- 3 main data points captured & related, person's portrait thumbnail, full name, organisation (diocese, religious order or SPCM)
- Roles of individuals – student, teacher, guest of honour (Bishop or lay staff)



Maynooth University National University of Ireland Maynooth | Arts and Humanities Institute

Issues faced

- Degradation of class pieces (environment, e.g. light damage, water damage & paper quality)
- Low resolution jpeg images
- Language (Irish/Latin) + typeface & Gaelic script
- Solution = manual entry



Data ingestion & data cleaning

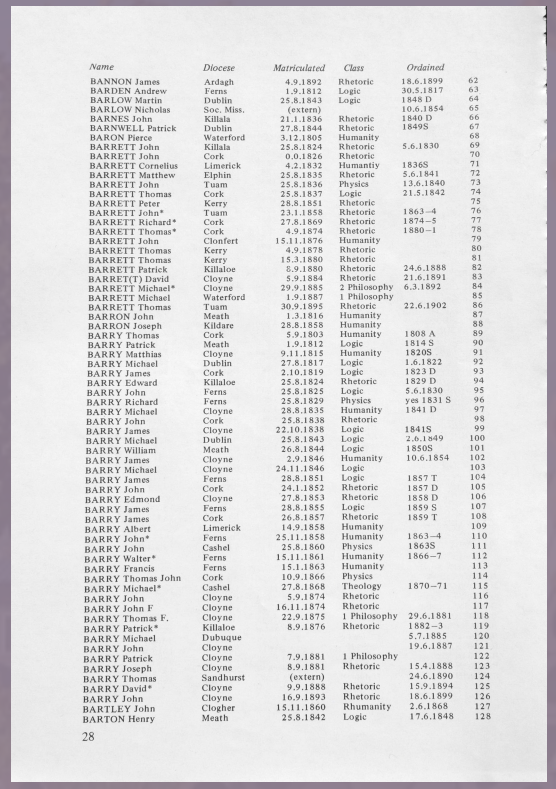
- Over 7,000 biographical entries added to the database
- Standardisation of organisation names (dioceses & religious orders)
- Merging of duplicate entries

Data enrichment

- Classpieces contain less than 50% of SPCM students since 1795
- We used the Hamell 2 list of students (1895-2002)
- Manual cross referencing Hamell 2 with database entries
- Translation & language issues (< 100 untranslated)
- Matriculation / ordination events added to entries
- Total = 3,000+ entries

Hamell list ingestion

- Photograph the primary source (book)
- prepare images for optical character recognition (OCR)
- extract content in spreadsheets (csv)
- verify results
- data transformations – normalizations
- mapping of csv data to data model
- automatically ingest content from spreadsheets
- verify results



Name	Diocese	Matriculated	Class	Ordained	
BANNON James	Ardagh	4.9.1892	Rhetoric	18.6.1899	62
BARDEN Andrew	Ferns	1.9.1812	Logic	30.5.1817	63
BARLOW Martin	Dublin	25.8.1843	Logic	1848 D	64
BARLOW Nicholas	Soc. Miss.	(extern)		10.6.1854	65
BARNES John	Killala	21.1.1836	Rhetoric	1840 D	66
BARNWELL Patrick	Dublin	27.8.1844	Rhetoric	1849S	67
BARON Pierce	Waterford	3.12.1805	Humanity		68
BARRETT John	Killala	25.8.1824	Rhetoric	5.6.1830	69
BARRETT John	Cork	1.1.1826	Rhetoric		70
BARRETT Cornelius	Limerick	4.2.1832	Humantiy	1836S	71
BARRETT Matthew	Elphin	25.8.1835	Rhetoric	5.6.1841	72
BARRETT John	Tuam	25.8.1836	Physics	13.6.1840	73
BARRETT Thomas	Cork	25.8.1837	Logic	21.5.1842	74
BARRETT Peter	Kerry	28.8.1851	Rhetoric		75
BARRETT John*	Tuam	23.1.1858	Rhetoric	1863–4	76
BARRETT Richard*	Cork	27.8.1869	Rhetoric	1874–5	77
BARRETT Thomas*	Cork	4.9.1874	Rhetoric	1880–1	78
BARRETT John	Clonfert	15.11.1876	Humanity		79
BARRETT Thomas	Kerry	4.9.1878	Rhetoric		80
BARRETT Thomas	Kerry	15.3.1880	Rhetoric		81
BARRETT Patrick	Killaloe	5.9.1880	Rhetoric	24.6.1888	82
BARRET(T) David	Cloyne	5.9.1884	Rhetoric	21.6.1891	83
BARRETT Michael*	Cloyne	29.9.1885	2 Philosophy	6.3.1892	84
BARRETT Michael	Waterford	1.9.1887	1 Philosophy		85
BARRETT Thomas	Tuam	30.9.1895	Rhetoric	22.6.1902	86
BARRON John	Meath	1.3.1816	Humanity		87
BARRON Joseph	Kildare	28.8.1858	Humanity		88
BARRY Thomas	Cork	5.9.1803	Humanity	1808 A	89
BARRY Patrick	Meath	1.9.1812	Logic	1814 S	90
BARRY Matthias	Cloyne	9.11.1815	Humanity	1820S	91
BARRY Michael	Dublin	27.8.1817	Logic	1.6.1822	92
BARRY James	Cork	2.10.1819	Logic	1823 D	93
BARRY Edward	Killaloe	25.8.1824	Rhetoric	1829 D	94
BARRY John	Ferns	25.8.1825	Logic	5.6.1830	95
BARRY Richard	Ferns	25.8.1829	Physics	yes 1831 S	96
BARRY Michael	Cloyne	28.8.1835	Humanity	1841 D	97
BARRY John	Cork	25.8.1838	Rhetoric		98
BARRY James	Cloyne	22.10.1838	Logic	1841S	99
BARRY Michael	Dublin	25.8.1843	Logic	2.6.1849	100
BARRY William	Meath	26.8.1844	Logic	1850S	101
BARRY James	Cloyne	2.9.1846	Humanity	10.6.1854	102
BARRY Michael	Cloyne	24.11.1846	Logic		103
BARRY James	Ferns	28.8.1851	Logic	1857 T	104
BARRY John	Cork	24.1.1852	Rhetoric	1857 D	105
BARRY Edmond	Cloyne	27.8.1853	Rhetoric	1858 D	106
BARRY James	Ferns	28.8.1855	Logic	1859 S	107
BARRY James	Cork	26.8.1857	Rhetoric	1859 T	108
BARRY Albert	Limerick	14.9.1858	Humanity		109
BARRY John*	Ferns	25.11.1858	Humanity	1863–4	110
BARRY John	Cashel	25.8.1860	Physics	1863S	111
BARRY Walter*	Ferns	15.11.1861	Humanity	1866–7	112
BARRY Francis	Ferns	15.1.1863	Humanity		113
BARRY Thomas John	Cork	10.9.1866	Physics		114
BARRY Michael*	Cashel	27.8.1868	Theology	1870–71	115
BARRY John	Cloyne	5.9.1874	Rhetoric		116
BARRY John F	Cloyne	16.11.1874	Rhetoric		117
BARRY Thomas F.	Cloyne	22.9.1875	1 Philosophy	29.6.1881	118
BARRY Patrick*	Killaloe	8.9.1876	Rhetoric	1882–3	119
BARRY Michael	Dubuque			5.7.1885	120
BARRY John	Cloyne			19.6.1887	121
BARRY Patrick	Cloyne	7.9.1881	1 Philosophy		122
BARRY Joseph	Cloyne	8.9.1881	Rhetoric	15.4.1888	123
BARRY Thomas	Sandhurst	(extern)		24.6.1890	124
BARRY David*	Cloyne	9.9.1888	Rhetoric	15.9.1894	125
BARRY John	Cloyne	16.9.1893	Rhetoric	18.6.1899	126
BARTLEY John	Clogher	15.11.1860	Rhumanity	2.6.1868	127
BARTON Henry	Meath	25.8.1842	Logic	17.6.1848	128

Name	Diocese	Matriculated	Class	Ordained	
BANNON James	Ardagh	4.9.1892	Rhetoric	18.6.1899	62
BARDEN Andrew	Ferns	1.9.1812	Logic	30.5.1817	63
BARLOW Martin	Dublin	25.8.1843	Logic	1848 D	64
BARLOW Nicholas	Soc. Miss.	(extern)		10.6.1854	65
BARNES John	Killala	21.1.1836	Rhetoric	1840 D	66
BARNWELL Patrick	Dublin	27.8.1844	Rhetoric	1849S	67
BARON Pierce	Waterford	3.12.1805	Humanity		68
BARRETT John	Killala	25.8.1824	Rhetoric	5.6.1830	69
BARRETT John	Cork	0.0.1826	Rhetoric		70
BARRETT Cornelius	Limerick	4.2.1832	Humantiy	1836S	71
BARRETT Matthew	Elphin	25.8.1835	Rhetoric	5.6.1841	72
BARRETT John	Tuam	25.8.1836	Physics	13.6.1840	73
BARRETT Thomas	Cork	25.8.1837	Logic	21.5.1842	74
BARRETT Peter	Kerry	28.8.1851	Rhetoric		75
BARRETT John*	Tuam	23.1.1858	Rhetoric	1863–4	76
BARRETT Richard*	Cork	27.8.1869	Rhetoric	1874–5	77
BARRETT Thomas*	Cork	4.9.1874	Rhetoric	1880–1	78
BARRETT John	Clonfert	15.11.1876	Humanity		79
BARRETT Thomas	Kerry	4.9.1878	Rhetoric		80
BARRETT Thomas	Kerry	15.3.1880	Rhetoric		81
BARRETT Patrick	Killaloe	5.9.1880	Rhetoric	24.6.1888	82
BARRET(T) David	Cloyne	5.9.1884	Rhetoric	21.6.1891	83
BARRETT Michael*	Cloyne	29.9.1885	2 Philosophy	6.3.1892	84
BARRETT Michael	Waterford	1.9.1887	1 Philosophy		85
BARRETT Thomas	Tuam	30.9.1895	Rhetoric	22.6.1902	86
BARRON John	Meath	1.3.1816	Humanity		87
BARRON Joseph	Kildare	28.8.1858	Humanity		88
BARRY Thomas	Cork	5.9.1803	Humanity	1808 A	89
BARRY Patrick	Meath	1.9.1812	Logic	1814 S	90
BARRY Matthias	Cloyne	9.11.1815	Humanity	1820S	91
BARRY Michael	Dublin	27.8.1817	Logic	1.6.1822	92
BARRY James	Cork	2.10.1819	Logic	1823 D	93
BARRY Edward	Killaloe	25.8.1824	Rhetoric	1829 D	94
BARRY John	Ferns	25.8.1825	Logic	5.6.1830	95
BARRY Richard	Ferns	25.8.1829	Physics	yes 1831 S	96
BARRY Michael	Cloyne	28.8.1835	Humanity	1841 D	97
BARRY John	Cork	25.8.1838	Rhetoric		98
BARRY James	Cloyne	22.10.1838	Logic	1841S	99
BARRY Michael	Dublin	25.8.1843	Logic	2.6.1849	100
BARRY William	Meath	26.8.1844	Logic	1850S	101
BARRY James	Cloyne	2.9.1846	Humanity	10.6.1854	102
BARRY Michael	Cloyne	24.11.1846	Logic		103
BARRY James	Ferns	28.8.1851	Logic	1857 T	104
BARRY John	Cork	24.1.1852	Rhetoric	1857 D	105
BARRY Edmond	Cloyne	27.8.1853	Rhetoric	1858 D	106
BARRY James	Ferns	28.8.1855	Logic	1859 S	107
BARRY James	Cork	26.8.1857	Rhetoric	1859 T	108
BARRY Albert	Limerick	14.9.1858	Humanity		109
BARRY John*	Ferns	25.11.1858	Humanity	1863–4	110
BARRY John	Cashel	25.8.1860	Physics	1863S	111
BARRY Walter*	Ferns	15.11.1861	Humanity	1866–7	112
BARRY Francis	Ferns	15.1.1863	Humanity		113
BARRY Thomas John	Cork	10.9.1866	Physics		114
BARRY Michael*	Cashel	27.8.1868	Theology	1870–71	115
BARRY John	Cloyne	5.9.1874	Rhetoric		116
BARRY John F	Cloyne	16.11.1874	Rhetoric		117
BARRY Thomas F.	Cloyne	22.9.1875	1 Philosophy	29.6.1881	118
BARRY Patrick*	Killaloe	8.9.1876	Rhetoric	1882–3	119
BARRY Michael	Dubuque			5.7.1885	120
BARRY John	Cloyne			19.6.1887	121
BARRY Patrick	Cloyne	7.9.1881	1 Philosophy		122
BARRY Joseph	Cloyne	8.9.1881	Rhetoric	15.4.1888	123
BARRY Thomas	Sandhurst	(extern)		24.6.1890	124
BARRY David*	Cloyne	9.9.1888	Rhetoric	15.9.1894	125
BARRY John	Cloyne	16.9.1893	Rhetoric	18.6.1899	126
BARTLEY John	Clogher	15.11.1860	Rhumanity	2.6.1868	127
BARTON Henry	Meath	25.8.1842	Logic	17.6.1848	128

28

Maynooth University National University of Ireland Maynooth | Arts and Humanities Institute

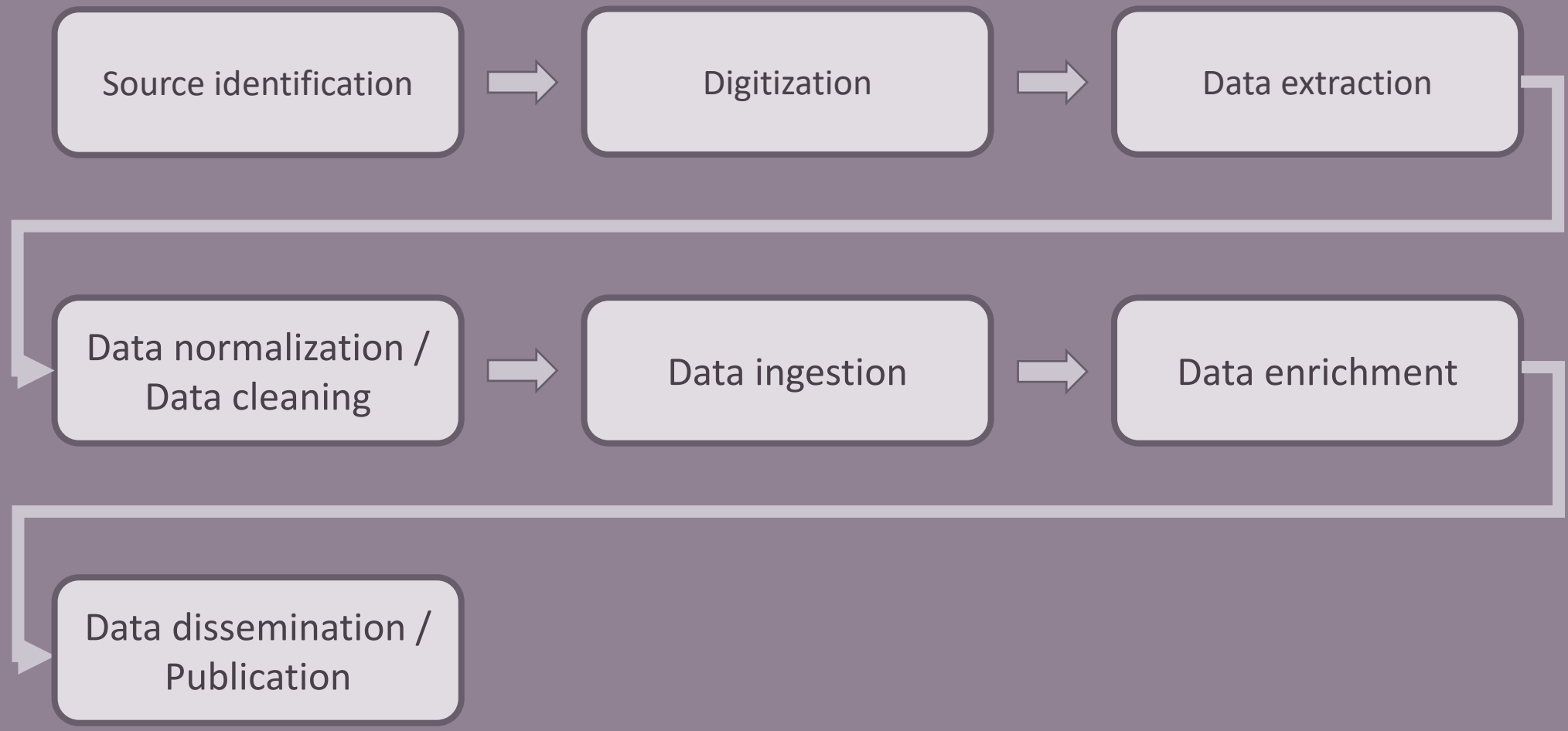
Content discussion

Technical WP

- Information workflow
- Data modelling
- Technical infrastructure development

Maynooth University
National University of Ireland Maynooth
Arts and Humanities Institute

Information workflow

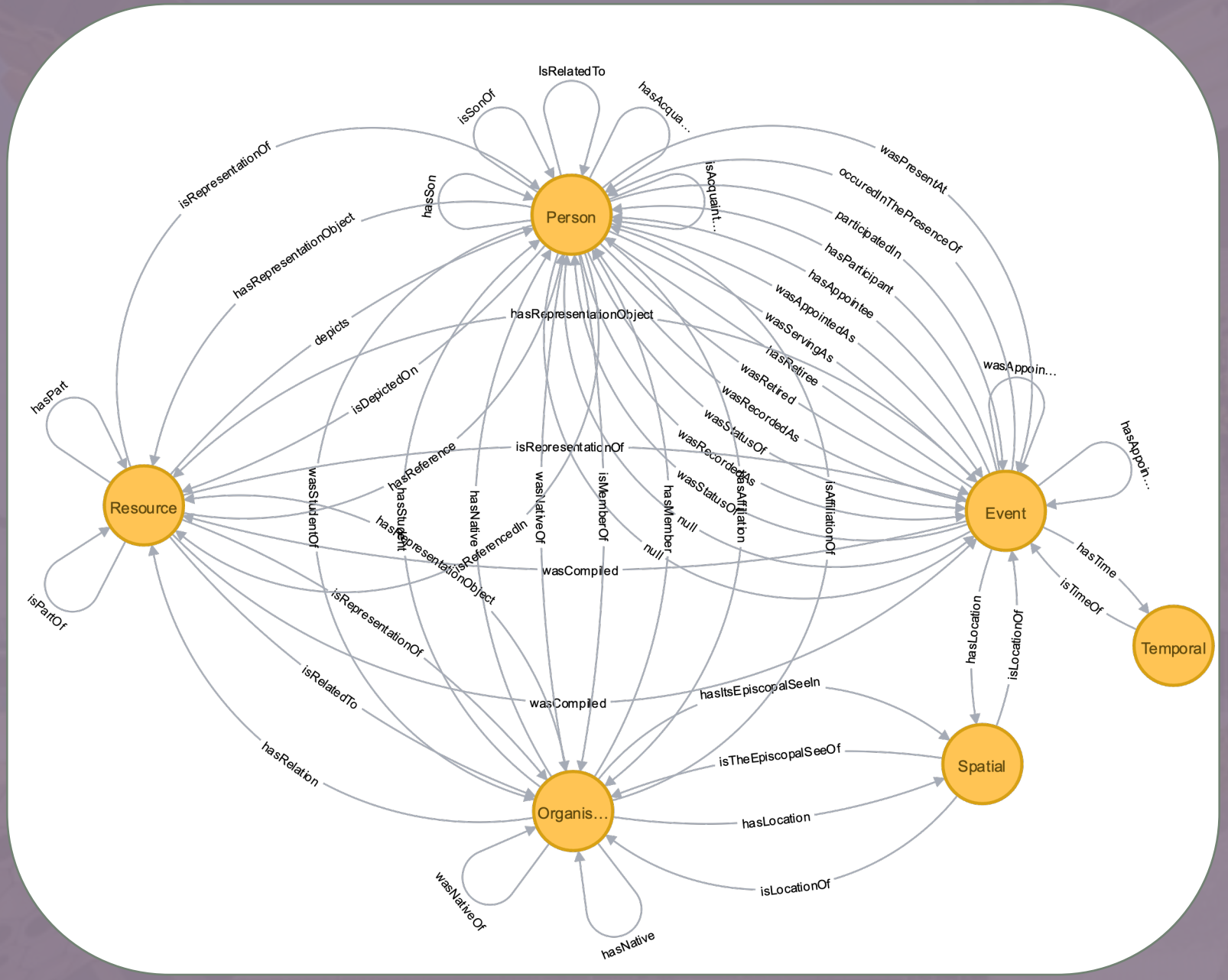


Data modelling

- Simplicity
- Flexibility
- Extensibility

Main entities

- Event
- Organisation
- Person
- Resource
- Spatial
- Temporal



Maynooth University
National University of Ireland Maynooth
Arts and Humanities Institute

Person attributes

- _id
- Label
- Honorific Prefix
- First name
- Middle name
- Last name
- First name soundex
- Last name soundex
- Alternate appelations
- Description
- Type
- Status
- Created by
- Created at
- Updated by
- Updated at

Maynooth University National University of Ireland Maynooth | Arts and Humanities Institute

Person relations

- participated in -> Event
- was appointed as -> Event
- was appointed to -> Event
- was awarded -> Event
- was present at -> Event
- was recorded as -> Event
- was retired -> Event
- was serving as -> Event

- has affiliation -> Organisation
- is member of -> Organisation
- was native of -> Organisation
- was student of -> Organisation
- has son -> Person
- is son of -> Person
- has representation object -> Resource
- is depicted on -> Resource
- is referenced in -> Resource

Event attributes

- _id
- Label
- Description
- Type
- Status
- Created by
- Created at
- Updated by
- Updated at

Event relations

- has relation -> Organisation
- has appointee -> Person
- has participant -> Person
- has retiree -> Person
- occurred in the presence of -> Person
- was awarded -> Person
- was held by -> Person
- was occupied by -> Person
- was status of -> Person
- was compiled -> Resource
- has representation object -> Resource
- has time -> Temporal
- has location -> Spatial

Resource attributes

- _id
- Label
- Alternate labels
- Description
- Filename
- Metadata
- Original location
- Paths
- Type
- System type
- Status
- Created by
- Created at
- Updated by
- Updated at

Maynooth University | Arts and Humanities Institute

Resource relations

- is representation of -> Event
- was compiled -> Event
- is related to -> Organisation
- is representation of -> Organisation
- depicts -> Person
- has part -> Resource
- has reference -> Person
- is part of -> Resource
- is representation of -> Person

Organisation attributes

- _id
- Label
- Label soundex
- Alternate labels
- Description
- Type
- Status
- Created by
- Created at
- Updated by
- Updated at

Maynooth University National University of Ireland Maynooth | Arts and Humanities Institute

Organisation relations

- is related to -> Event
- has member -> Person
- has student -> Person
- has native -> Person
- is affiliation of -> Person
- has relation -> Resource
- has representation object -> Resource
- has location -> Spatial
- has its episcopal see in -> Spatial

Spatial attributes

- _id
- Label
- Street address
- Locality
- Region
- Postal code
- Country
- Latitude
- Longitude

- Type
- Note
- Rawdata
- Created by
- Created at
- Updated by
- Updated at

Maynooth University
National University of Ireland Maynooth
Arts and Humanities Institute

Spatial relations

- is the episcopal see of -> Organisation
- is location of -> Organisation
- is location of -> Event

Temporal attributes

- _id
- Label
- Start date
- End date
- Format
- Created by
- Created at
- Updated by
- Updated at

Temporal relations

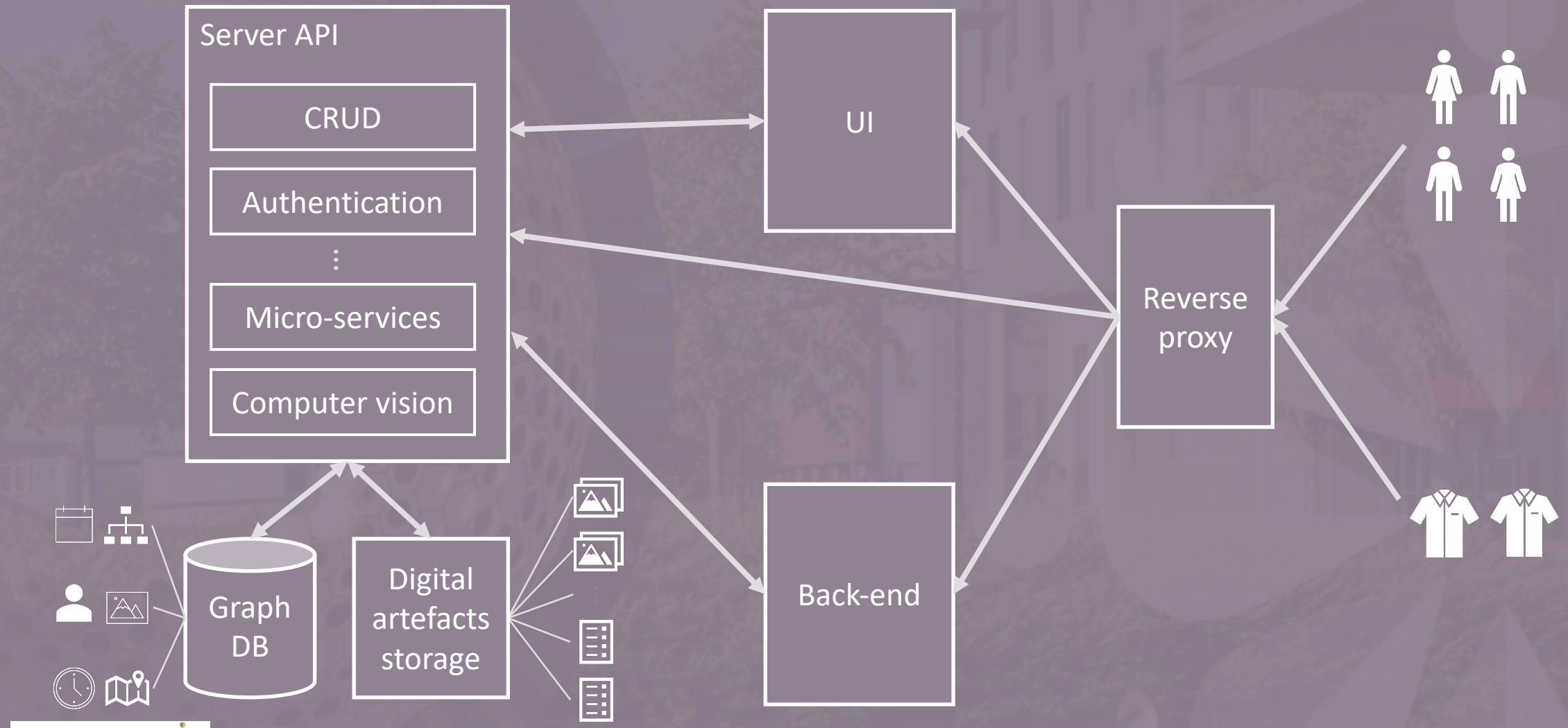
- is time of -> Event

Taxonomies

- Event type
- Matriculation class
- Ordination ranks
- Organisation type
- People roles
- Person type
- Relations type
- Resource system type
- User group

Data modelling discussion

Technical infrastructure



Maynooth University National University of Ireland Maynooth | Arts and Humanities Institute

Technical overview

The technical infrastructure of the Clericus project can be divided into three main components:

- Server API
- Content Management Admin Interface
- Content Delivery Public Interface

Server API (1/2)

- a collection of tools and services that enables agents (e.g. the front/back end applications) to communicate with the database and perform a series of tasks.
- written in **Javascript**, in particular **node** with **express**, a popular combination for building web applications.
- the main function of the Server API is that of a **REST API**.
- **API endpoints** (web URIs that invoke a certain function) have been developed to handle all the necessary **CRUD** operations on our data.
- All API endpoints are documented with the use of **apiDoc** for easier communication between the development team.

Server API (2/2)

- Data validation and sanitisation is being handled by the Server API to ensure the database data integrity with classes defined for each main database entity and methods to ensure data manipulation occurs without mistakes.
- Role based user authentication layer (**argon2**, **passport, JSON web Token)**
- A set of custom tools has been developed to support various operations necessary for the system, e.g.:
  - tool for the **AI** supported automatic identification of people and features on classpieces
  - dynamic query builder
  - various command line tools developed per case to perform the automatic ingestion of a new data collection.

Content Management Admin Interface

The Content Management Admin Interface is a user interface build with **React**, a Javascript framework. It is available only to authenticated users, users that possess the necessary rights to update the content of the project.

CMAI features

- admin access to browse and manipulate data
- separate customised views for each distinct data type
- custom image viewer
- annotation tool
- data model editor
- taxonomy editor
- content management
- custom tools, e.g.
  - classpiece importer
  - query builder
- user management

Maynooth University National University of Ireland Maynooth | Arts and Humanities Institute

Content Delivery Public Interface

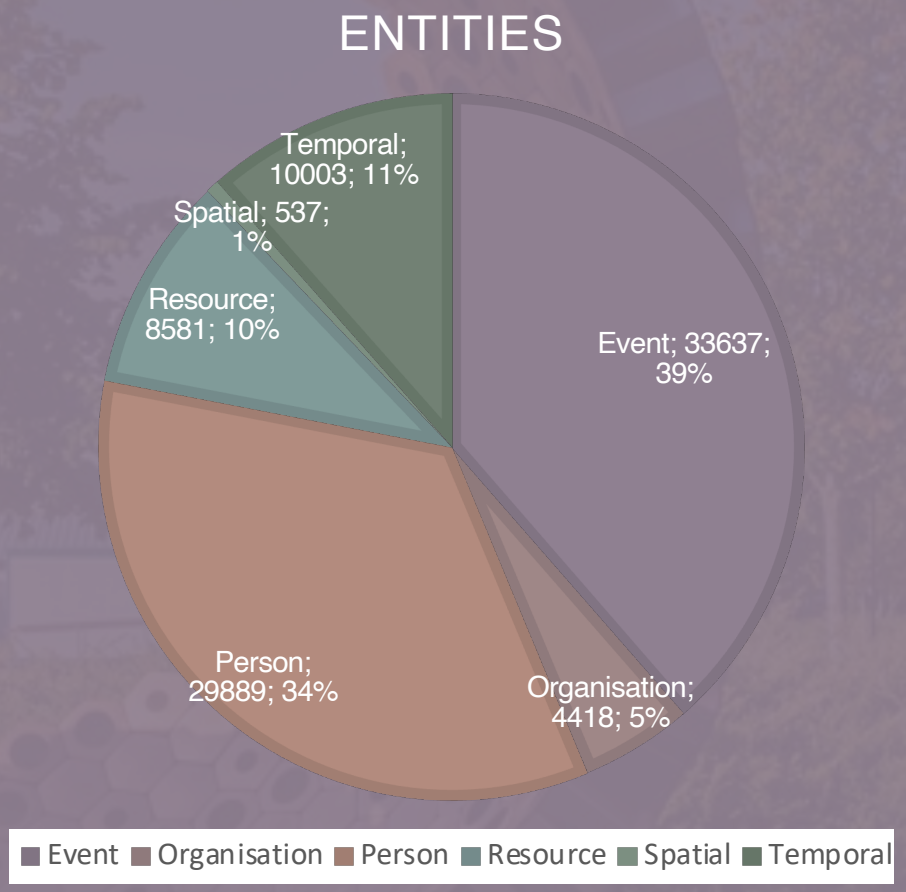
The Content Delivery Public Interface is a user interface build with **React**, a Javascript framework. **Bootstrap** is used as the main templating toolkit to ensure a modern, responsive layout. The **D3** library is being employed for the graph data visualisations, expanded with **PixiJS** for better performance and more efficient rendering. **Leaflet** is the javascript library employed for making the responsive map visualisations. Finally **Sass**, an extension to the **CSS** stylesheet language, is being employed to add modern and efficient styling to the application.

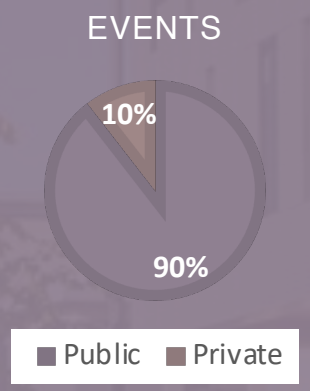
CDPI features

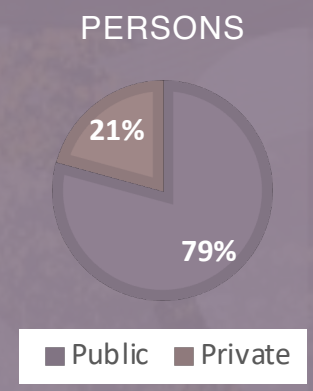
- Home screen with custom components
  - content carousel
  - quick statistics
  - content highlights…
- Articles list and page views
- Content entities distinct filtered list views and entry level views
- Classpieces grid view
- custom classpiece viewer with search support
- data visualisations for aggregated data and distinct entries
  - heatmap
  - timeline
  - graph network

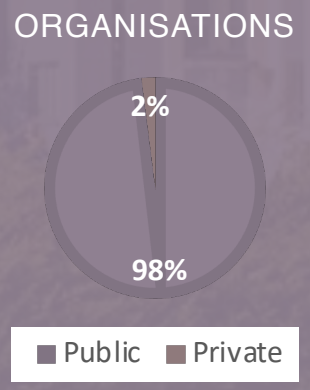
Technical infrastructure discussion

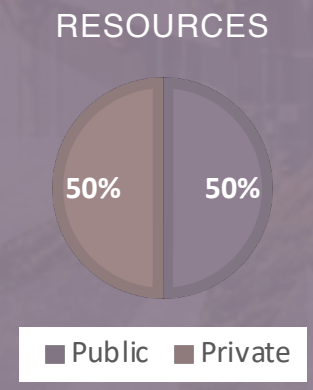
Statistics - entities





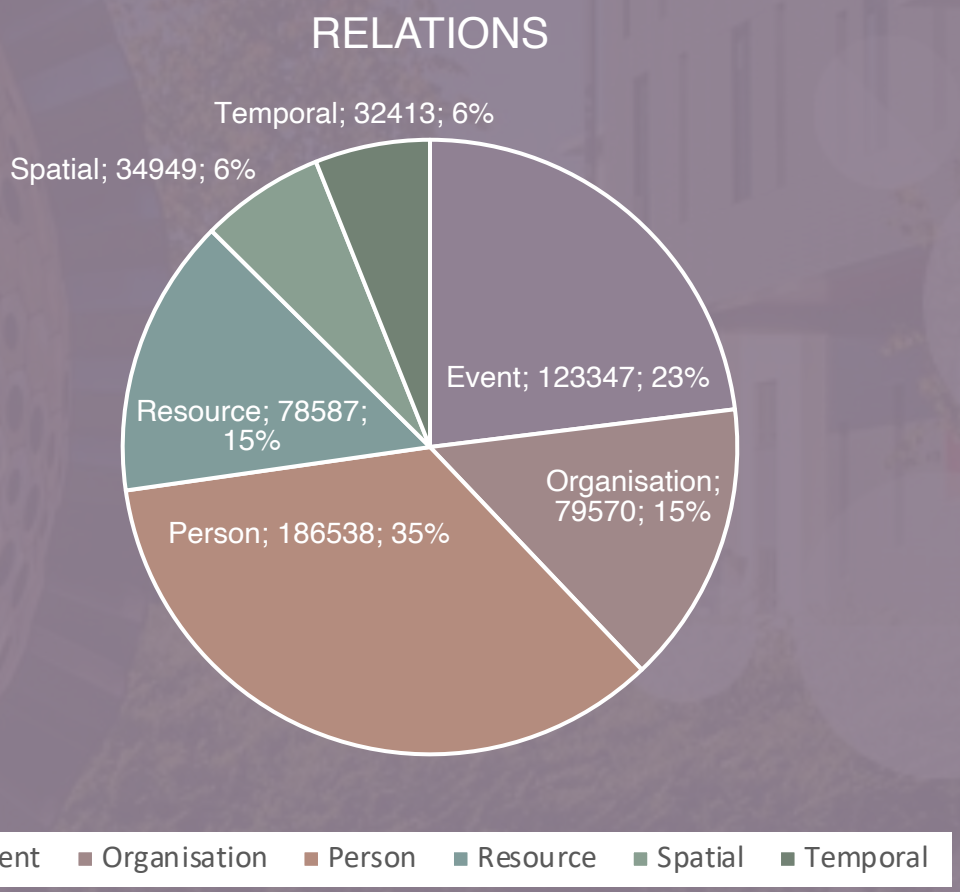






Maynooth University National University of Ireland Maynooth | Arts and Humanities Institute

Statistics - relations



Future work

- Dynamic data import tool
- Data analysis WP
- Dissemination WP
- Communications WP
- Optimisation of graph network
- ….

Maynooth University National University of Ireland Maynooth | Arts and Humanities Institute

Find out more at

https://clericus.ie

Maynooth University
National University of Ireland Maynooth
Arts and Humanities Institute