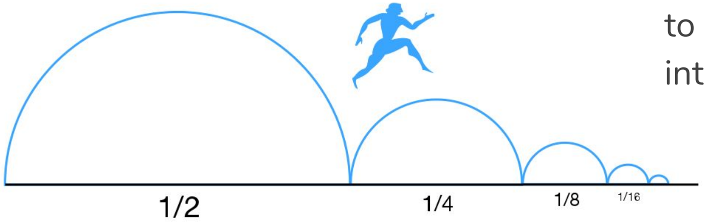
Data for History & The Semantic Data Challenge in Historical Research

EADH 2018, Galway, Ireland
9/12/2018

A Background Paradox

Digital documentation tools allow historians to record **ever more** and varied **information** in information systems **in analytical detail**, supporting their ability to analyze primary **evidence** in order to provide evidence to **historical argumentation**.



The increased capability to record such data, **without** accompanying **formal** information **modelling** training, **data standards** and the **infrastructure** required to host, maintain and provide access to such analytic data **means** that this **information** largely remains **difficult** to find, hard to interpret and not interoperable.

The dream

What if we were able to develop and provide a sufficiently rich set of tools and techniques for data modelling for historians that would empower them to generate truly FAIR data [Findable - Accessible - Interoperable - Reusable] meaning:

- Analytic Research results would be encoded using both standard schema and data values
- Analytic Research results could be accessed and reused on the fly
- An ever growing web of rich research data in compatible form would provide a massive graph of provenanced, historical data which would stand as a basic research tool to historians

The sceptical moment

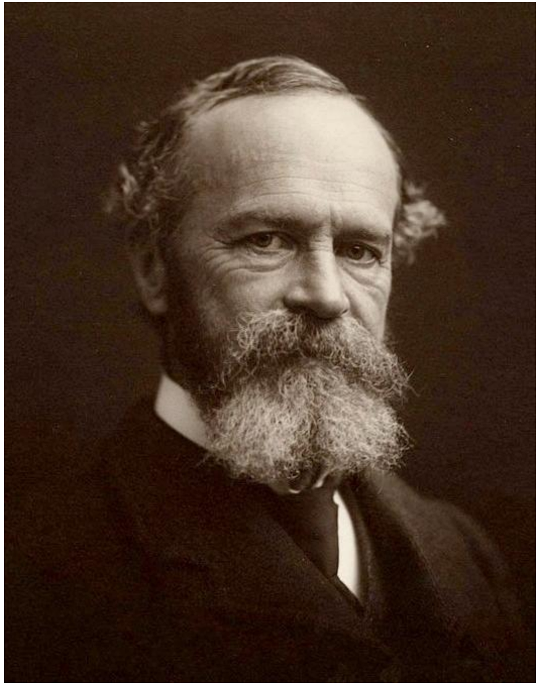
Historical data is massively rich and diverse, covering any aspect of human activity and its effects through time, by what possible means could such standardization be achieved?

- Would such standardization not simply stifle the expressivity of research?
- Would such standardization not unnecessarily slow research?
- How could any imagined standard model manage the many potential epistemic diagreements, shifts and ruptures?

The pragmatist rejoinder

While historical research is premised on disagreement in search of the truth, it grounds this disagreement in an analysis of facts and sources:

- It is the representation of facts, their sources, and their interrelation for which we can aim to offer a general model
- Strides forward in formal ontological modelling and practices in maintaining consistent reference data makes this possibility ever more imaginable
- The work of creating analytic, historical data is already a basic task of research; it is worth trying to propose a general model for it



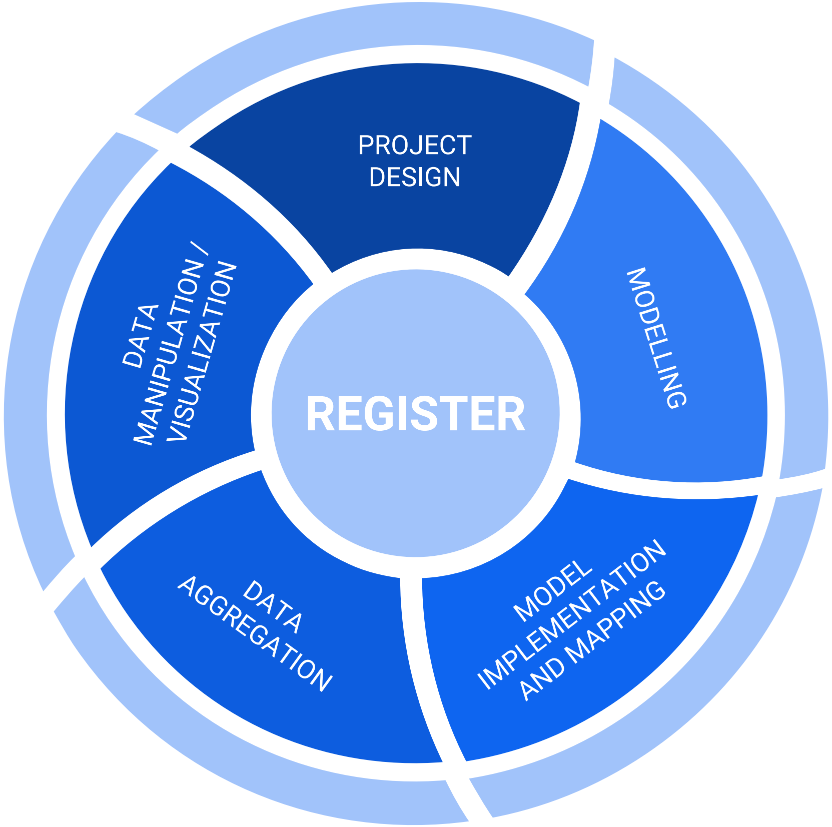
Getting there: Data for History

- Historians from across Europe interested in:
    - Generating a common semantic model
    - Adopting CIDOC CRM and adapting experience from Symogi
    - Using well known reference resources for standardizing
    - Creating and sharing semantic data
    - Applying semantic data tools for historical research

Data for History Consortium

Membership	42
Membership Type	Historians, Computer Scientists, Digital Humanists
Country Representation	Austria, France, Germany, Greece, Italy, Luxemburg, Netherlands, Switzerland, United Kingdom
Institution Type Representation	Memory Institutions, Research Institutes, Universities
Founded	24/11/2017 @ Lyon, France
Steering Committee	Francesco Beretta (CNRS, LARHRA) George Bruseker (FORTH-ICS, CCI)

SEMANTIC DATA WORKFLOW MANAGEMENT


PROJECT DESIGN
MODELLING
MODEL IMPLEMENTATION AND MAPPING
DATA AGGREGATION
DATA MANIPULATION / VISUALIZATION
REGISTER

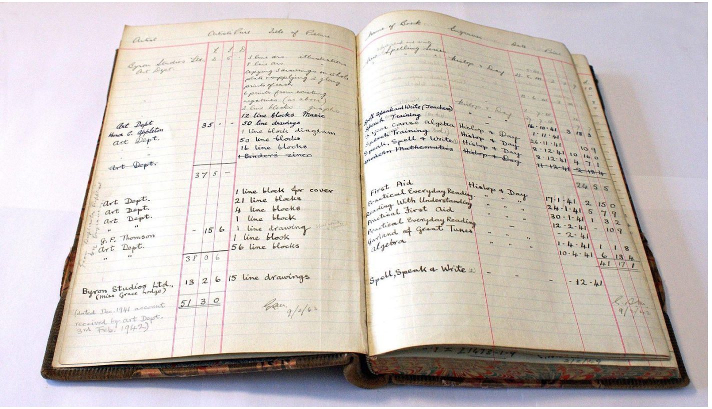
REGISTER

WHAT:

A database listing the resources available to the researcher/institution/consortium and their present state, which is updated as semantic project results are generated.

WHY:

We need to know what the present state of affairs is, what's on offer, what data is there, what state it is in, what partners are available, what resources can be accessed and we need to track what has been done, based on what.



https://commons.wikimedia.org/wiki/File:Ledger_detailing_external_work_commissioned_at_Holmes_McDougall_(4268190563).jpg

Tools Required	● Register
Tool Examples	1. Parthenos Entities and Architecture 2. Data for History Drupal

PROJECT DESIGN

WHAT:

Working on the basis of the known present state of affairs, project design in terms of existing dataset selection (if any), software toolkit selection, partner selection, research question design

WHY:

Proper Research Methodology informed by knowledge of latest state of affairs in resource base.

https://www.maxpixel.net/Pen-And-Paper-Notebook-Business-Office-Pen-Paper-1042595

Tools Required	● Pen ● Paper ● Education ● Project Management Software
Tool Examples	1. Bic 2. Moleskin 3. Not taking sides 4. RedMine

Modelling (Selection/Extension)

WHAT:

Selection of an adequate conceptual model to the needs of the particular project. The model may be fully adequate already or require extensions to an existing model in order to be realized. Can also document best practices for model.

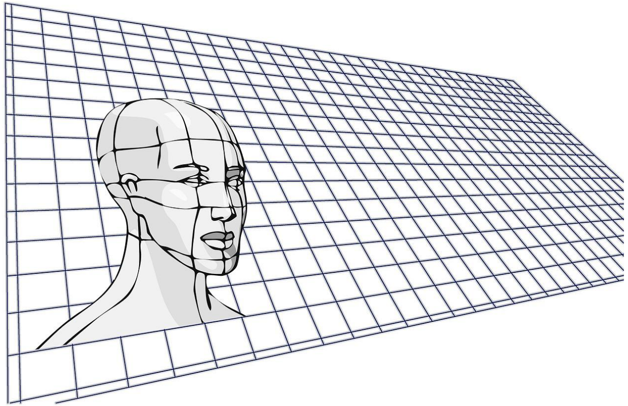
WHY:

Ensure adequate semantic representation for the field of research in order to answer questions, establish conditions of interoperability

https://www.flickr.com/photos/inpivic/5205915907

Tools Required	● ontology development management environment ● application profile designer ● Standardized vocabulary sources
Tool Examples	1. OntoMe 2. BBTalk
Output Types	● Ontologies (RDF/OWL) ● Application Profiles

Modelling Implementation/Mapping



https://pixabay.com/en/network-head-face-grid-human-516333/

WHAT:

For data production, a data curation environment should be chosen which can implement the model.

For pre-existing data, a data mapping environment should be chosen which can transform data to the model.

WHY:

Ensure new project data conformant to model while reusing and valorizing existing datasets.

Tools Required	● Data Curation Environment ● Data Mapping Environment
Tool Examples	1. ResearchSpace, Wisski, Arches, Qoqnus, Themas 2. 3M, Karma
Output Types	● Model Compatible RDF

Data Aggregation & Enrichment Processes

WHAT:

Once all desired sources are expressed in a conformant format (at schema and data level), semantic data should be ported to a common knowledge base for common use and exploration. This an also entail data alignment and enrichment.

WHY:

Ensure reproducible workflow of aggregation to common environment and provenance of semantic data.

https://commons.wikimedia.org/wiki/File:Flickr_-_ggallice_-_Caterpillar_aggregation.jpg

Tools Required	● Data Aligning/ Cleaning Tools ● Aggregation Management Environment
Tool Examples	1. Open Refine 2. DNet

Data Manipulation & Visualization

WHAT:

Use and interpret aggregate data to generate new knowledge

WHY:

Creating new knowledge is the final cause of all these processes.

https://www.flickr.com/photos/eric-delcroix/24519077825

Tools Required	● Knowledge graph search and visualization platform
Tool Examples	1. ResearchSpace

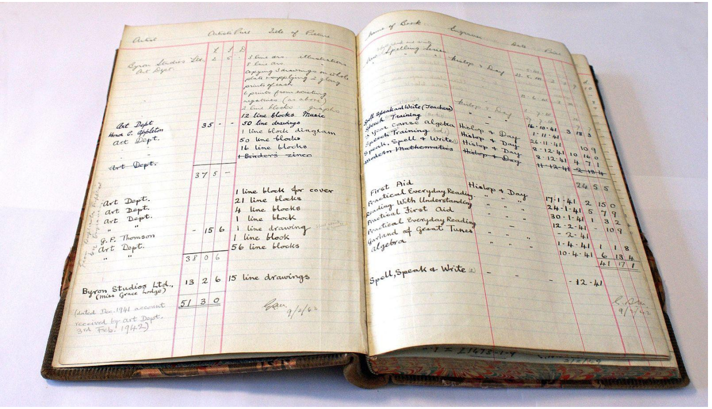
REGISTER

WHAT:

A database listing the resources available to the researcher/institution/consortium and their present state, which is updated as semantic project results are generated.

WHY:

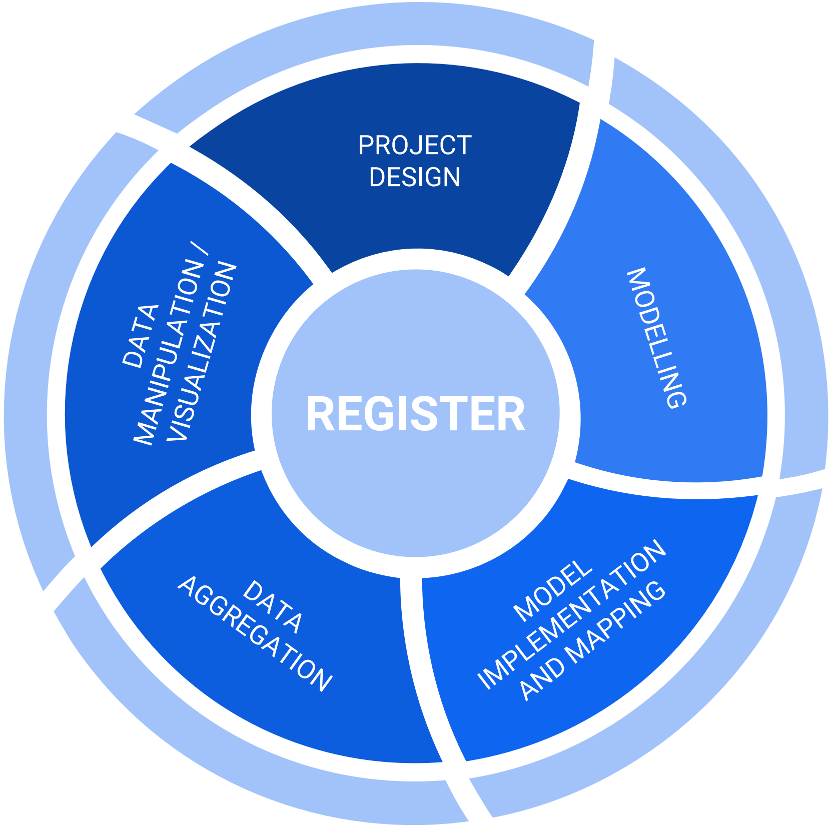
We need to know what the present state of affairs is, what's on offer, what data is there, what state it is in, what partners are available, what resources can be accessed and we need to track what has been done, based on what.



https://commons.wikimedia.org/wiki/File:Ledger_detailing_external_work_commissioned_at_Holmes_McDougall_(4268190563).jpg

Tools Required	● Register
Tool Examples	1. Parthenos Entities and Architecture 2. Specific Project Register e.g.: Data for History Drupal

SEMANTIC DATA WORKFLOW MANAGEMENT



PROJECT DESIGN
MODELLING
MODEL IMPLEMENTATION AND MAPPING
DATA AGGREGATION
DATA MANIPULATION / VISUALIZATION
REGISTER

More than a day's work

https://fr.wikipedia.org/wiki/Fichier:Maquette_de_Rome.JPG

Presentations

- A standard model for a prosopography of religious orders [B. Hours]
- APIS: Mapping the Austrian Biographic Dictionary to CIDOC CRM [M. Schlögl]
- Best practices for making data generated in automated linkage procedures readily re-usable [L. Petram]
- Building a domain specific Research Ontology from external Databases of Academic History [Edgard Marx]
- OntoMe : an ontology management environment for extending the CIDOC CRM to historical research sub-domains [F. Beretta]