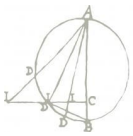
Digital Research Infrastructure for Digital Humanities

Building a Central Knowledge Graph for Research Projects

Robert Casties, Steffen Hennicke, Kim Pham
2022 ELAG



MPIWG
MAX PLANCK INSTITUTE
FOR THE HISTORY OF SCIENCE

Welcome everyone! And thanks for having us today!

We're going to talk about our Digital Research Infrastructure for Digital Humanities (DRIH), a project of our institute's library.

I will start with a couple of introductory remarks to give some context, Kim then will introduce the infrastructure in more detail, and Robert will talk more about the data modelling aspects.

The original goal of this project was to come up with a solution of what to do with the many digitized materials and digital research outputs of the different digital humanities projects at our institute.

MAX PLANCK INSTITUTE
FOR THE HISTORY OF SCIENCE

Max Planck Institute for the History of Science

Our institute, the Max Planck Institute for the History of Science, located in Berlin, currently has two active departments and, all together, we have about 350 researchers, many of them on rotating appointments.

Many of our projects involve digitization efforts of various kinds of historical source materials at the request of our researchers. These projects investigate many different aspects of the History of Science, by studying, for example, ancient cuneiform tablets, manuscripts from Early modern period Renaissance, Eurasian artefacts describing the world's astral history, chinese newspapers, gazetteers, or modern government documents.

This means that lots of digital data are being generated from these projects over a short period of time.

Over time, we amassed a large body of data that forms a comparatively long history of digital scholarship at the institute.

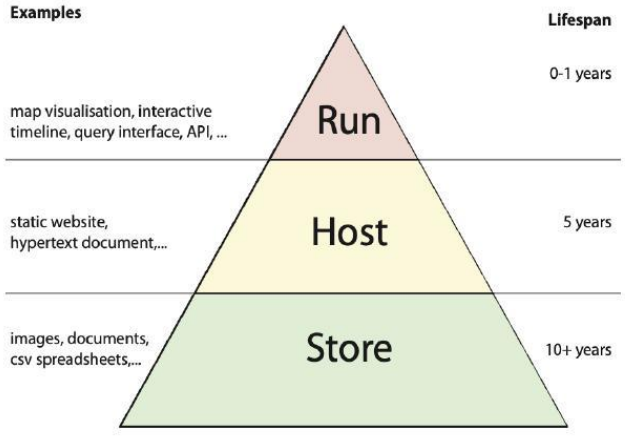
These data include by now over 5000 digitized rare books, alongside annotated documents, bibliographies, virtual exhibits, oral histories, and many publications.

Many of these data were originally put onto individual project sites operated by different staff and members at the institute. Today, we are still hosting over 130 sites from the last decade, even if the project has been completed.

Management of these sites, however, has become difficult, for reasons you might be familiar with, such as outdated technology, inoperable data standards that are difficult to migrate to newer systems, and a lack of consistent development practices for future sustainability.

MAX PLANCK INSTITUTE
FOR THE HISTORY OF SCIENCE

Digital Preservation



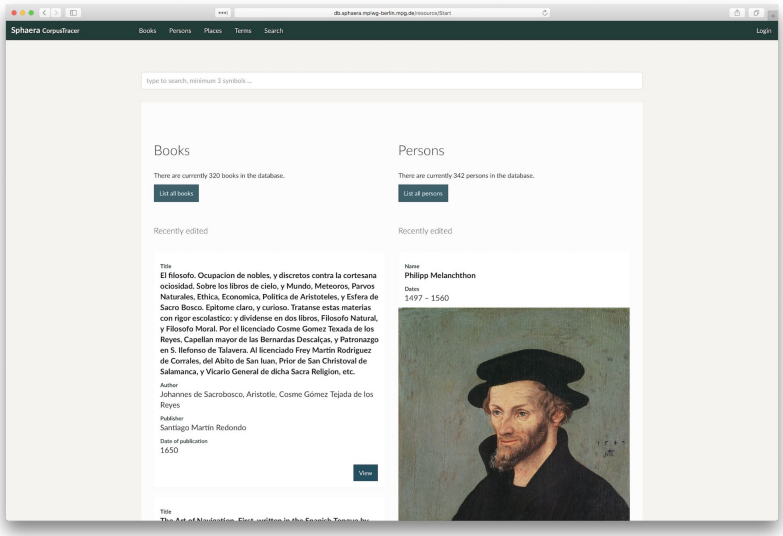
Taken from: Kräutli, Florian (2021): *Digital Research Afterlife: Graceful Degradation in Digital Humanities Projects*. 2021. http://www.kraeutli.com/index.php/2021/11/10/digital-research-afterlife-graceful-degradation-in-digital-humanities-projects/ (10.11.2021)

The solution for us was to think about how to consolidate these materials over time, moving away from practices that sort of grew organically to just keep the various sites as is, even if the project had become inactive or had been scattered across different servers.

Nowadays we are still building individual databases and unique websites for many projects, but now each project has to develop a research data management plan that describes the lifecycle for their digital products and how they will be handled by Research IT at the end of the project. Essentially that means coming up with a plan how to preserve the project's data.

Sphaera

MAX PLANCK INSTITUTE
FOR THE HISTORY OF SCIENCE



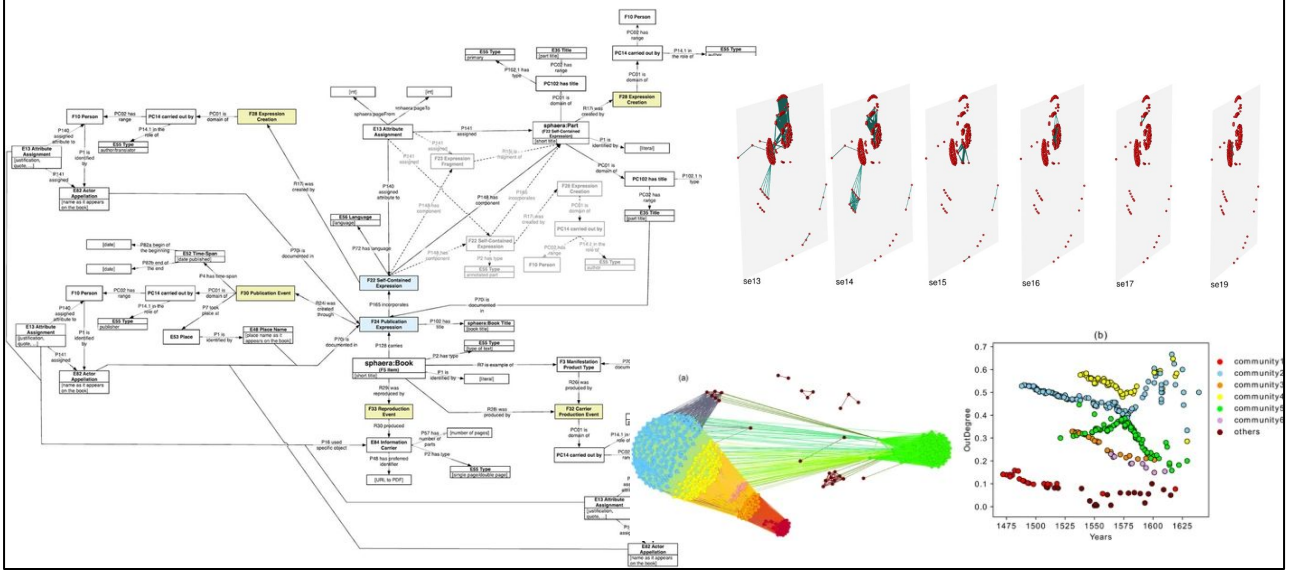
As a starting point to think about what a central digital humanities infrastructure could look like, we selected a model project, which would also be our first candidate project for the infrastructure, which is called Sphaera.

To us it was a model example on how the institute wishes to develop modern digital humanities projects, using open source tools and data standards that are suitable for the modern Web.

Sphaera investigates the knowledge tradition in early modern Europe that is interwoven with the history of one book: De Sphaera by Johannes de Sacrobosco. The text introduces a geocentric cosmology of the universe, so there is a correlation between the publication history and distribution of the text and the spread of this theory across Europe. The Sphaera database records bibliographic data of 320 digitized books published between 1472 and 1650, its geographic distribution, along with associated printers, publishers, and authors.

MAX PLANCK INSTITUTE
FOR THE HISTORY OF SCIENCE

Data modelling with linked data and CIDOC CRM

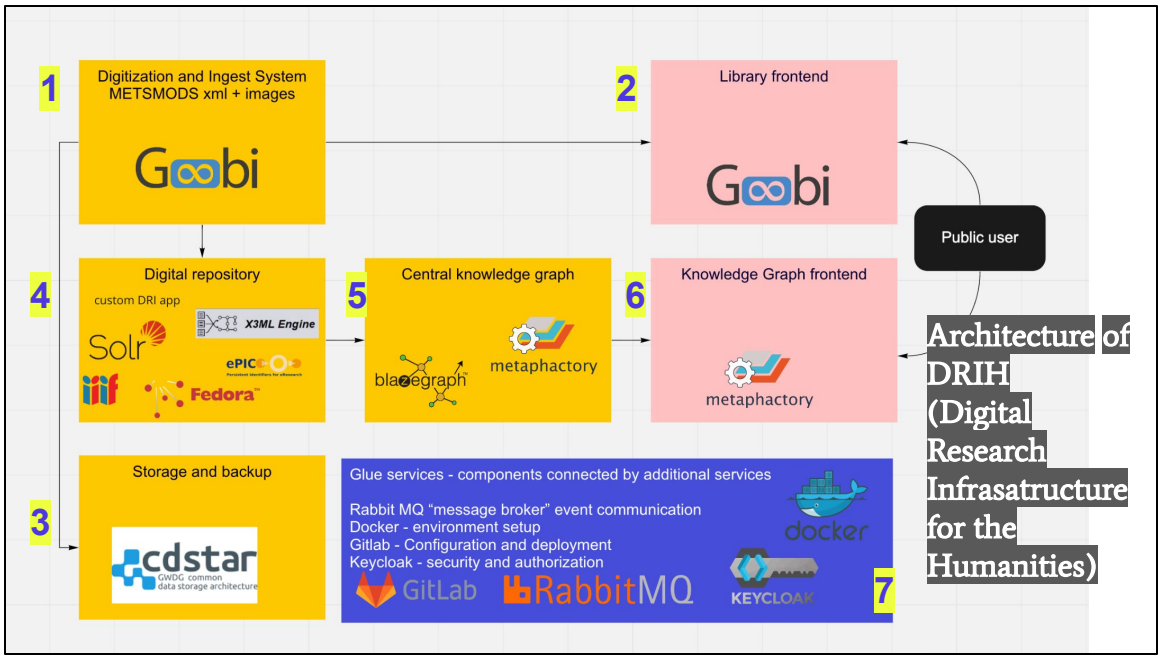


The project structures its data as CIDOC CRM and Linked Data. This was done first and foremost from a research perspective since the desire was to link individual entities and ideas along with additional links to resources in order identify and analyse the formation of the related social, historical and epistemic networks.

But it was also beneficial from a technical perspective to use CIDOC CRM and Linked Data since it removed the dependency on data to software in order for it to be usable.

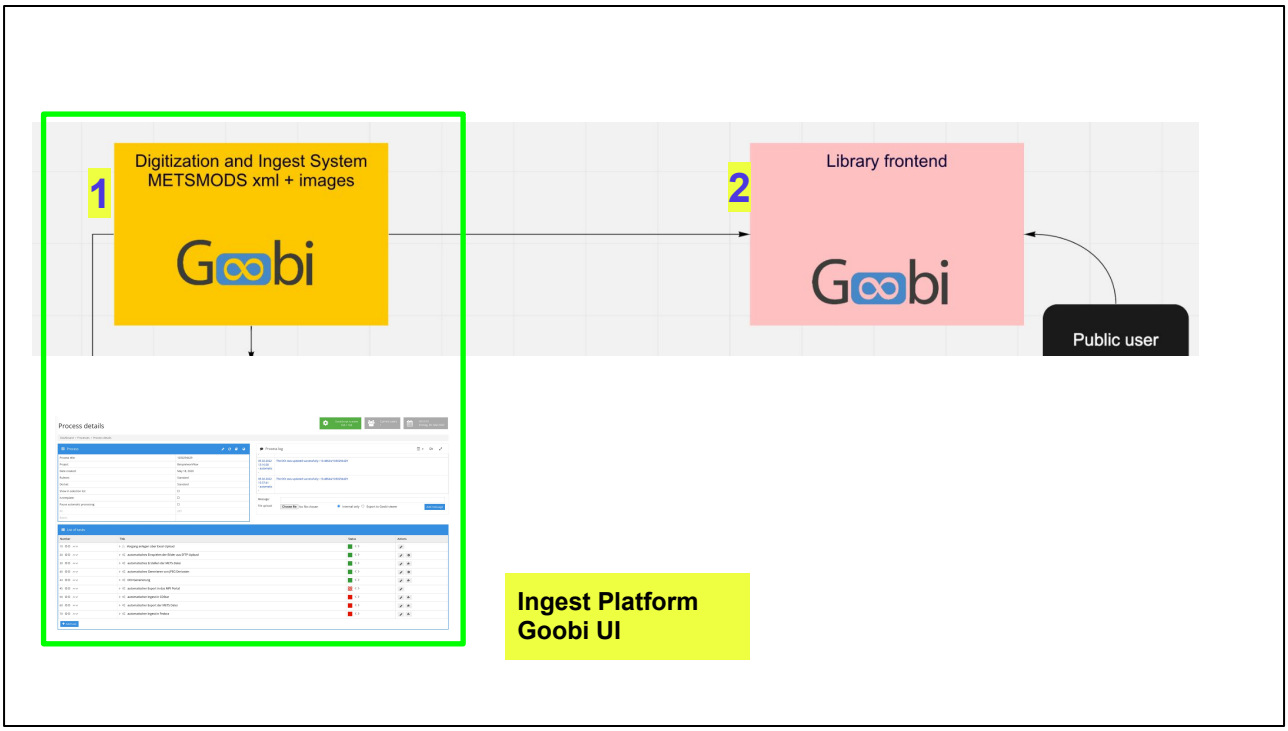
(The image on the left shows how the Sphaera project was modelled to describe the texts, authors, text parts, and additional publication information. The images on the right show that, based on this data, network analysis that was done to identify behaviours in the networks of knowledge.)

Kim will now introduce how this project helped us to inform the development of our Infrastructure for a more sustainable research data management at the institute.



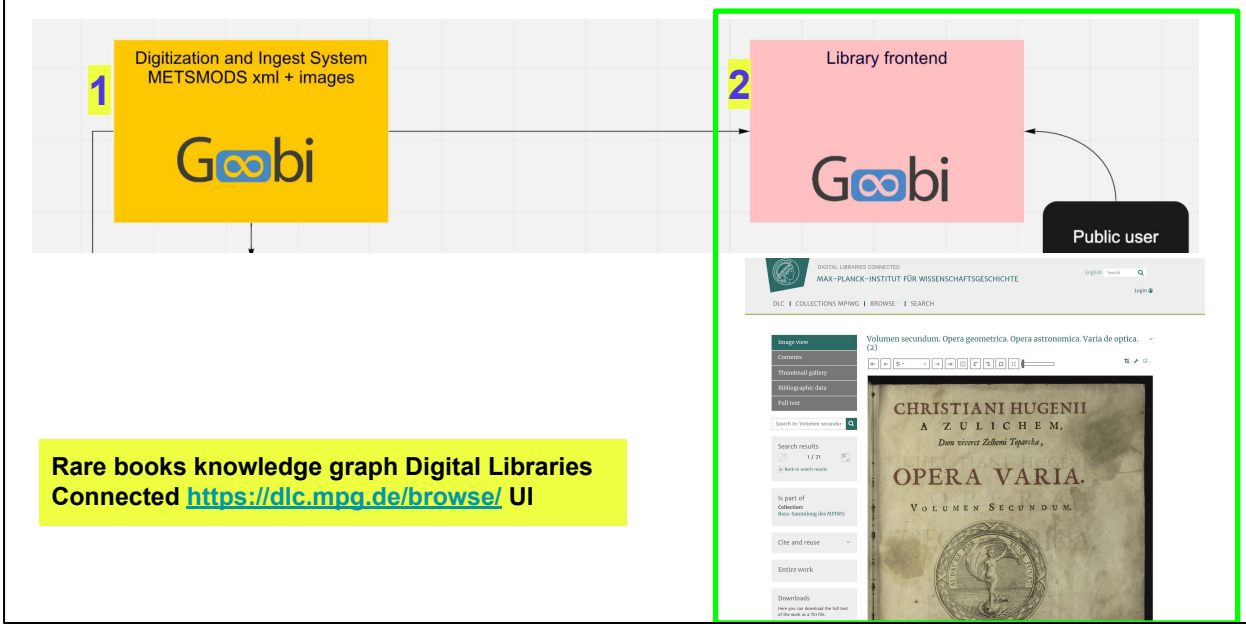
Thinking about how a project like Sphaera can be accessed in the long term, separate from the original project site, but still preserving its data and links to other materials, we developed a design which we refer to as DRIH - which stands for the digital research infrastructure for the humanities.

In order to select our technologies we looked at our existing systems that are being supported at the institute and at other similar research institutes, libraries, and academic institutions. What we settled on was an infrastructure to manage the digitization, and ingest of historical materials, that are stored and archived in a repository with additional bibliographic information, which is then linked to project data such as annotations, and information about entities such as people, places, and descriptions of things in a central knowledge graph.



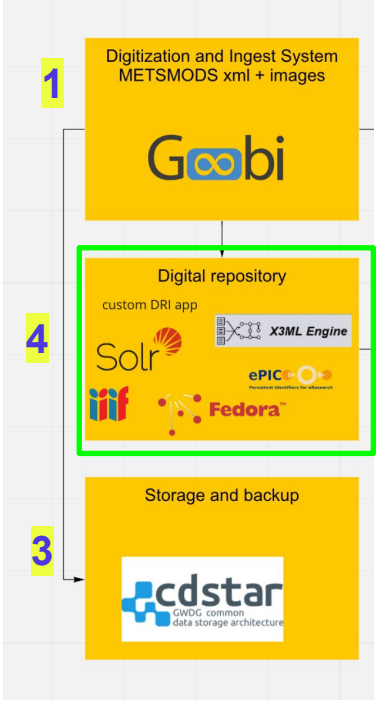
https://dlc.mpg.de/image/86733827X/5/LOG_0003/

- The first step in the DRIH workflow Starts with the digitization process.
- The library handles requests for digitization of rare books and texts using our Goobi system. Metadata is created using a template form that is stored as METS/MODS XML, which gets packaged with the scanned materials. Scanned materials are usually a collection of image files but they can also be links to the resource if it is available elsewhere or if we are not able to retain a digital copy
- This package is considered to be the digital object or digital representation of a scanned resource, and in this system we can see how it passes through the rest of the digital research infrastructure workflow into the other connected systems

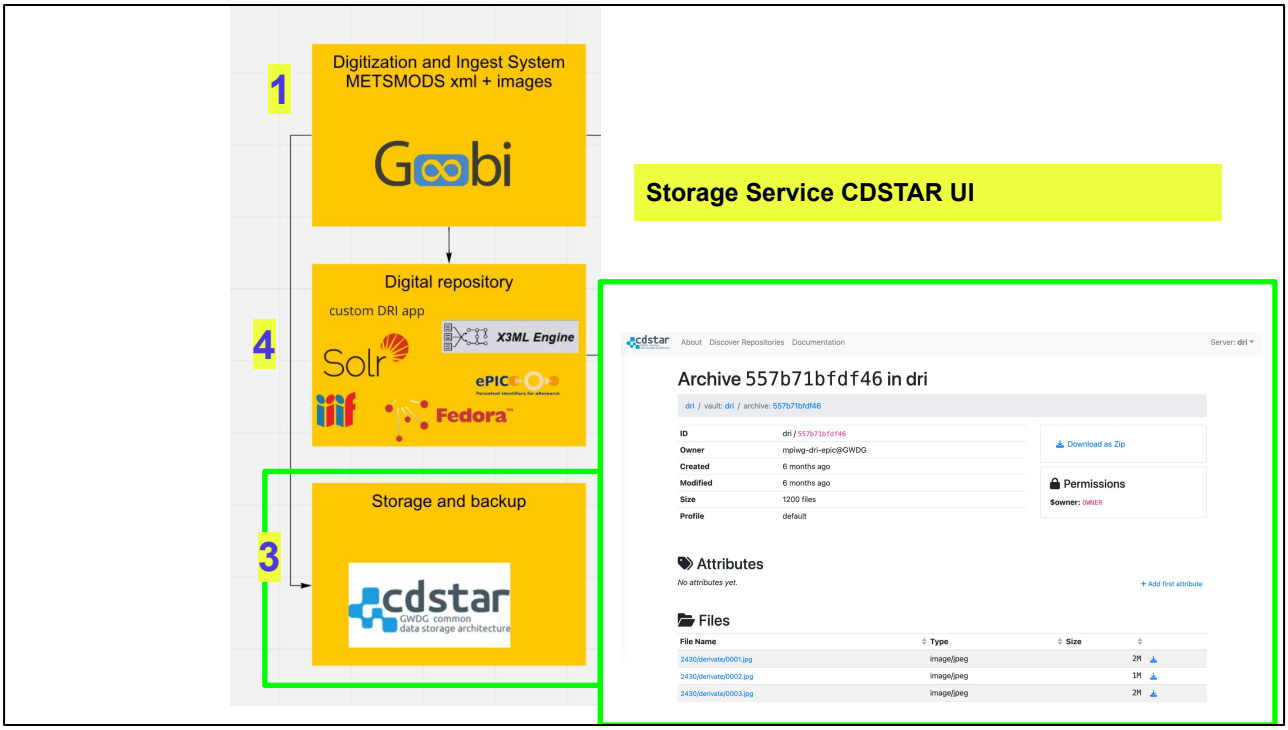


Rare books knowledge graph Digital Libraries Connected https://dlc.mpg.de/browse/ UI

- After this package gets ingested into Goobi, it is made public through the library's digital collections interface, called Digital Libraries Connected. This interface is shared with 3 other Max Planck institutions (art history, human development, and legal history and legal theory) so that users have a common interface to search materials from.

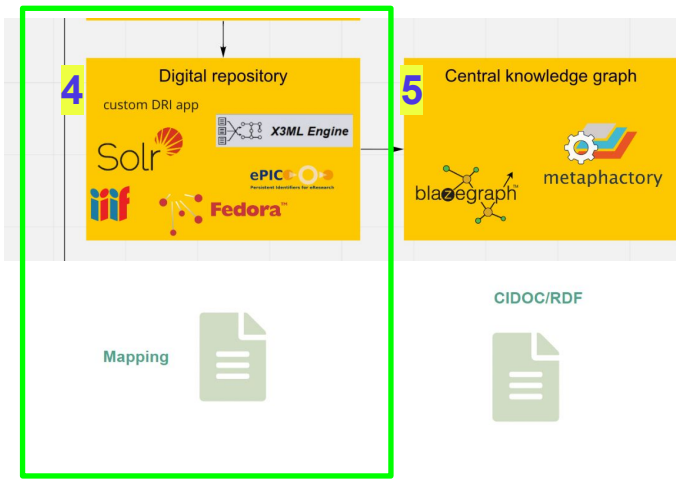


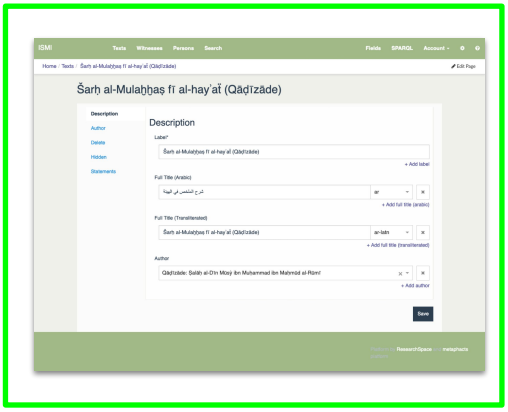
At the same time that is made viewable in the collections interface, the package contents are described in Fedora is used to track individual objects, their metadata and their contents in a systematic way.



- The actual image assets get stored in CDSTAR, which supports the preservation of these image assets by running constant checks to see if the data is still intact and not corrupted, creating backups, and generating technical information on the images.

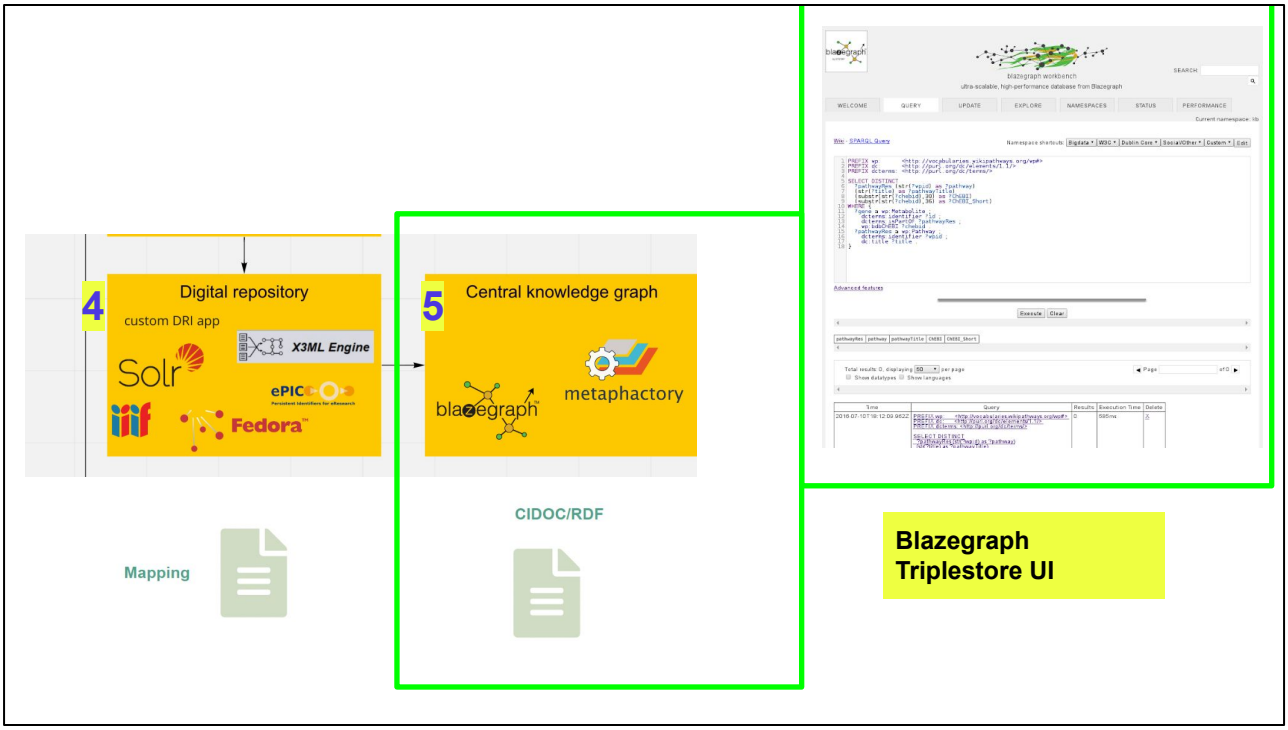
- We developed a digital repository that uses data coming from packages stored Fedora,The repository also is built with IIIF and Solr which is a search engine to index our content. The repository UI is also the site where we can access links of resources to data located in the central knowledge graph



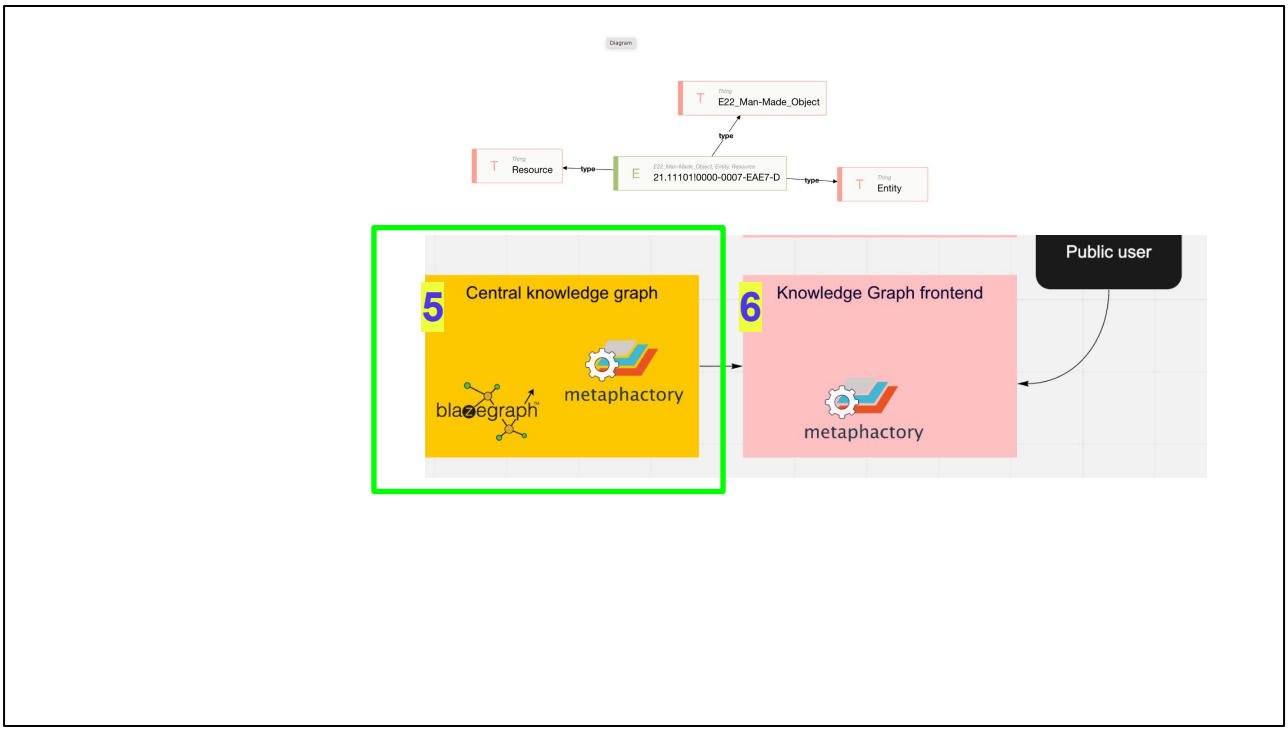


X3ML Mapping UI

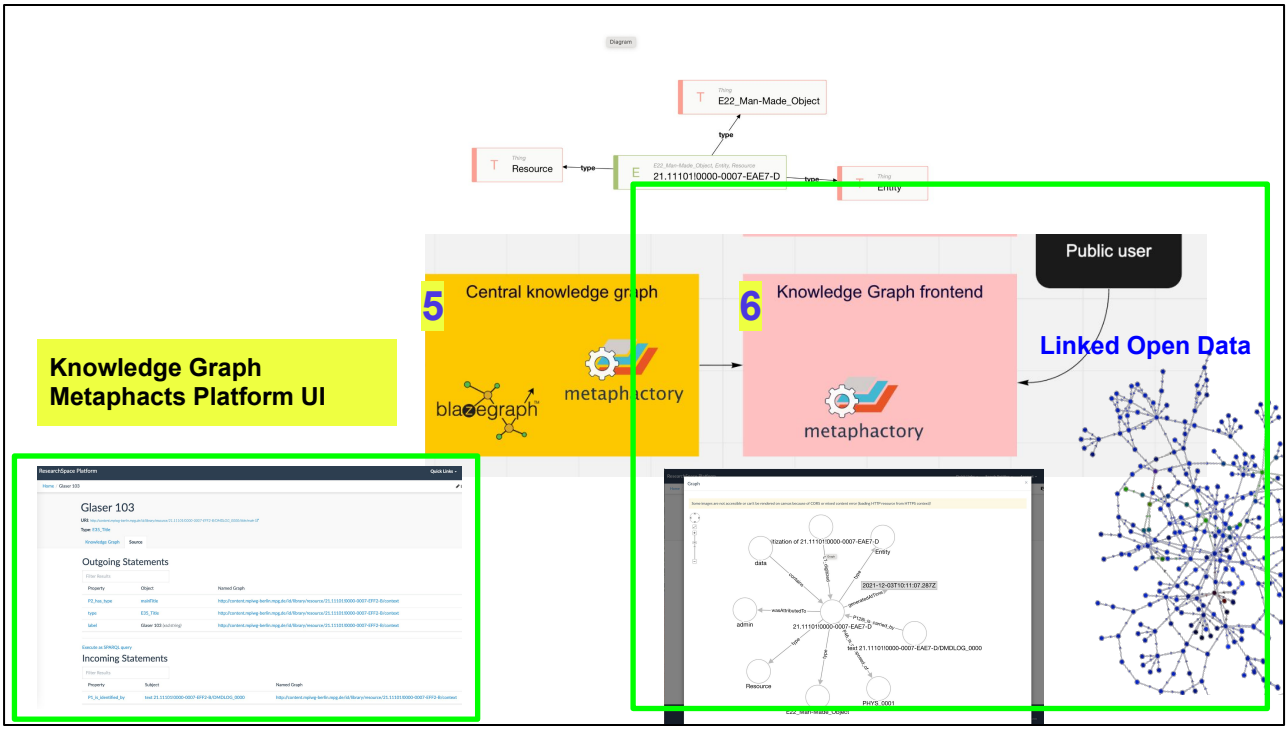
For the data to go eventually end up in the central knowledge graph, we have a tool called X3ML that helps map our digitized data packages into our schema implementation of linked data and CIDOC CRM.



The tool does the conversion and then passes the data into our knowledge graph environment, ResearchSpace also known as Metaphacts, which is supported by Blazegraph.



Our central knowledge graph system contains data coming from Goobi, which is the digitized rare books, but at the same time we are also inputting directly into the knowledge graph generated research data if it does not fall into the realm of bibliographic data.

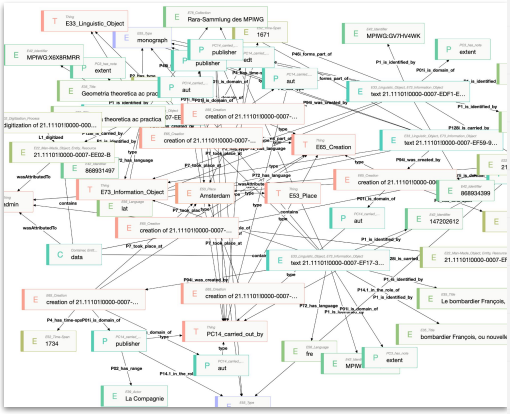


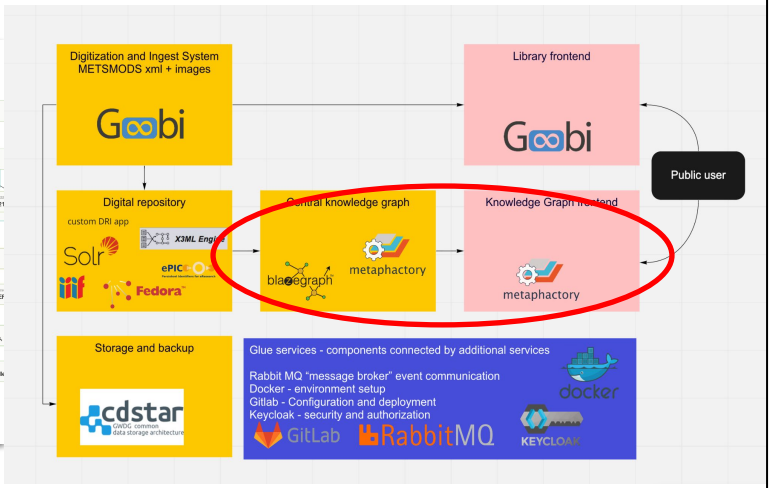
- This research data is ingested using the ResearchSpace API if it is already structured like linked data. If it isn't it then we have to spend some time to model and transform this data into the standards we use, case by case each project we have to work on individually to model and map data.
-

- I should mention that these all of these systems don’t all naturally work together, and that To integrate these systems we have to use smaller services in order for data to be passed on and stored in the different systems and for them to communicate with one another.
- We use RabbitMQ to pass on information about actions that are happening on in the system, transform information into formats that each system understands.
- We have to also manage the setup of these environments to work together on servers using Gitlab,
- and make sure these systems are secured and accessible only to the desired users using KEycloak.

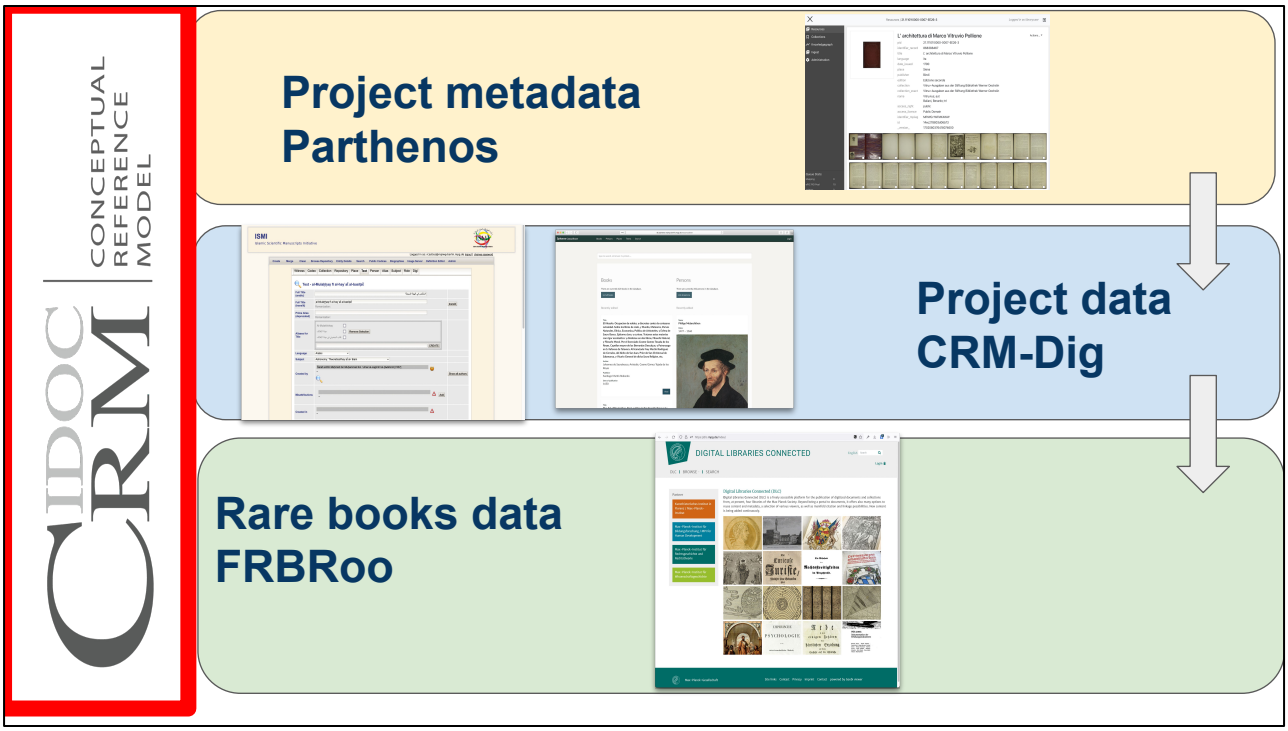
MAX PLANCK INSTITUTE FOR THE HISTORY OF SCIENCE

A Central Knowledge Graph (CKG) - Integrating project research data





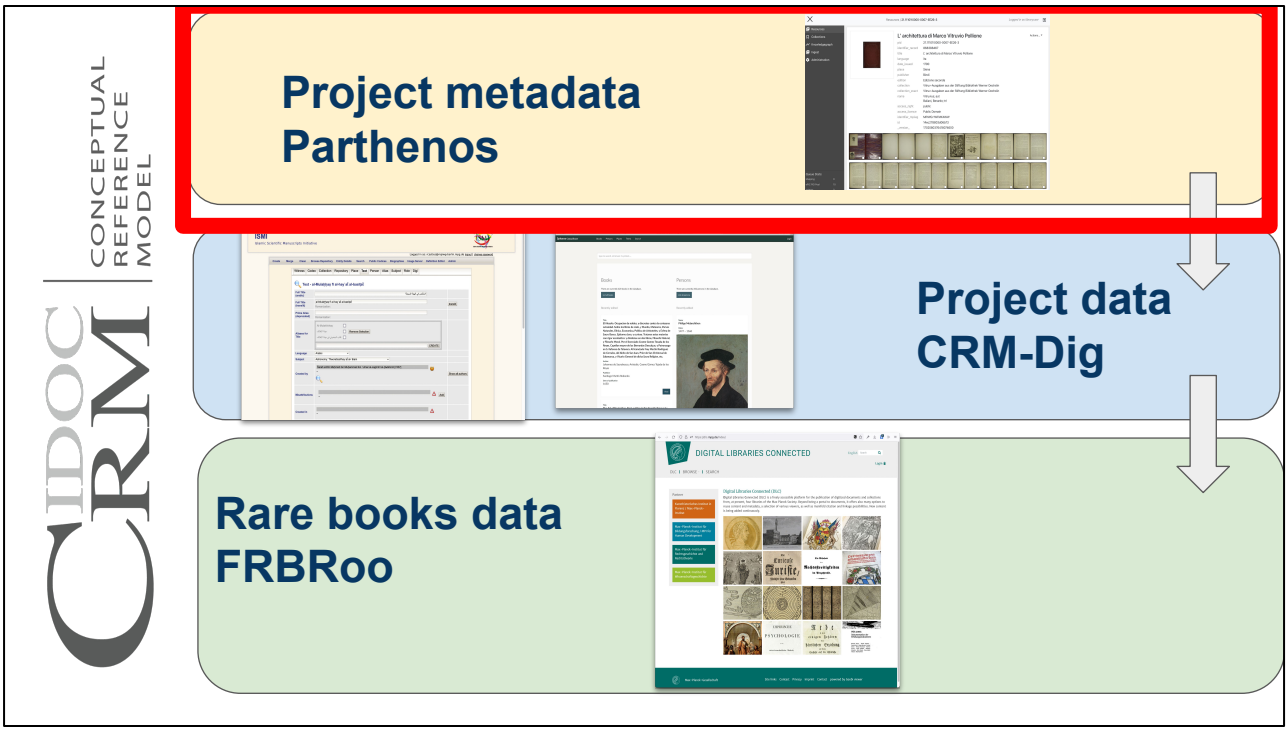
The central knowledge graph is considered the heart of this network of systems. It is meant to be the central starting point to query the sum of all project research data across project boundaries, or at least as much as we can integrate that data. To get project data to be queryable and discoverable through the central knowledge graph is part of our effort to keep our institute's research data as sustainable and reusable datasets that are integrated into the wider infrastructure of the institute.



To integrate our research data we had to use a common standard, in our case we used CIDOC CRM linked data as our common semantic target model as was already done by the Sphaera project.

It is widely used by cultural heritage institutions, and it has therefore a large active community working on similar projects in the area of cultural heritage and humanities research. CIDOC CRM allows us to express the complexity and intricacies contained in the various datasets that we get from the institute's projects as well as the ability to model cross-project generalities.This means that it allows for specialised use of its classes and properties, so that project data may define specific meanings while still retaining interoperability at the higher level of the ontology.

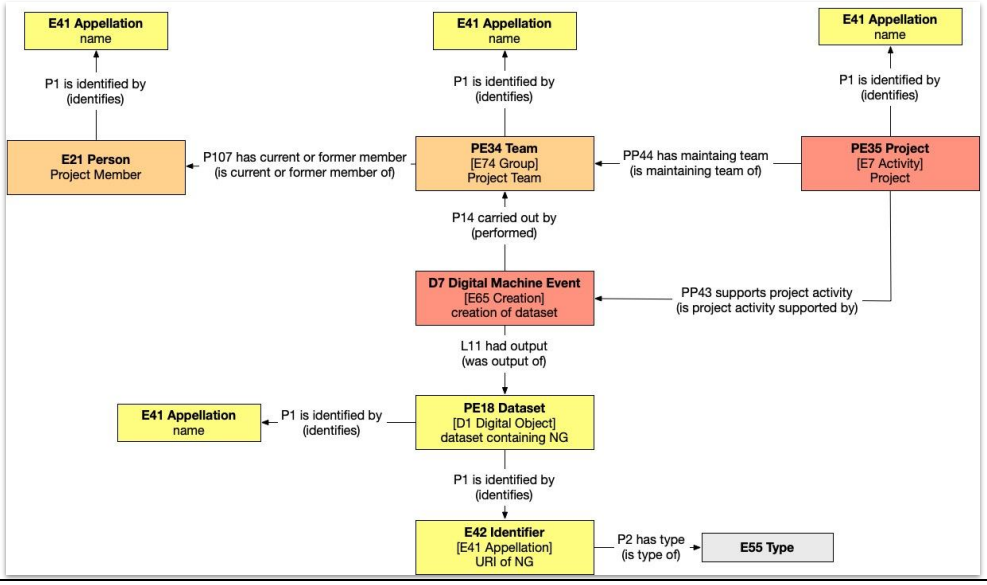
We also use other extensions of CIDOC-CRM to model particular aspects of our data.



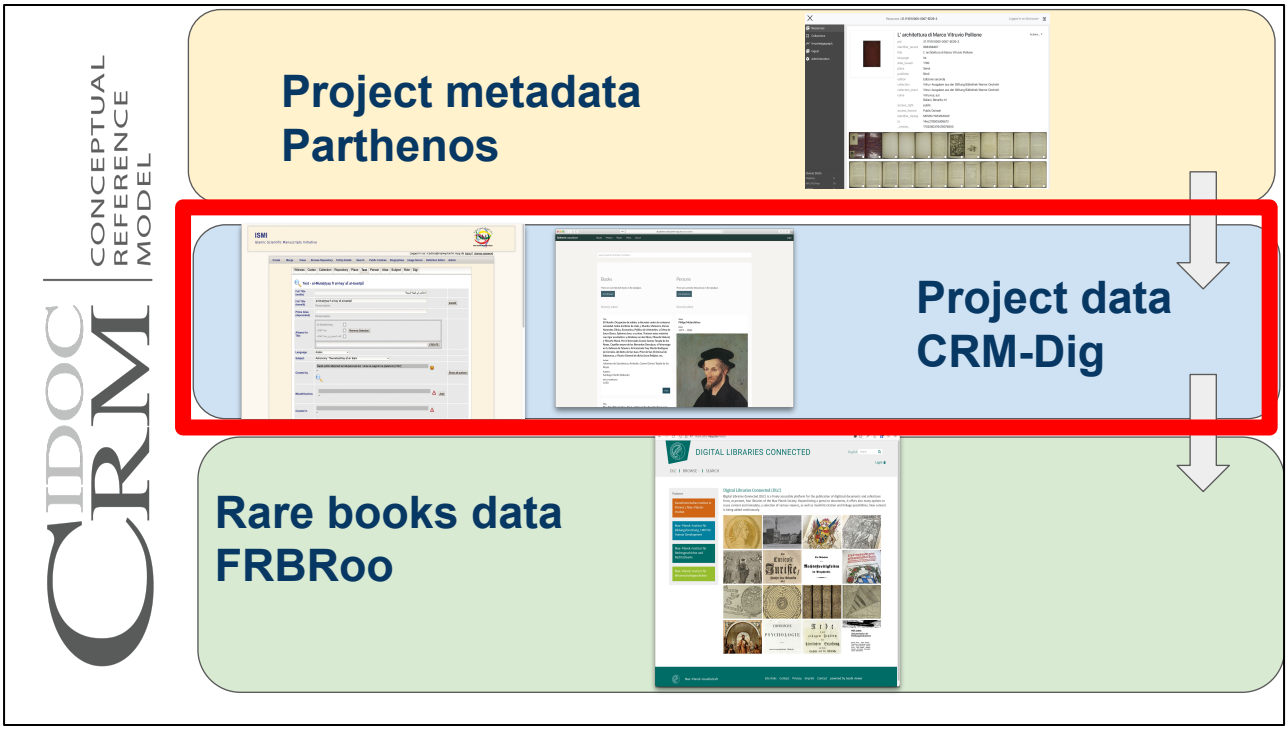
We use the Parthenos ontology in order to model the projects' organisational and historical context and the provenance of its digital artefacts. Parthenos allows us to add a high-level descriptive layer on top of the specific project datasets in the Knowledge Graph.

MAX PLANCK INSTITUTE FOR THE HISTORY OF SCIENCE

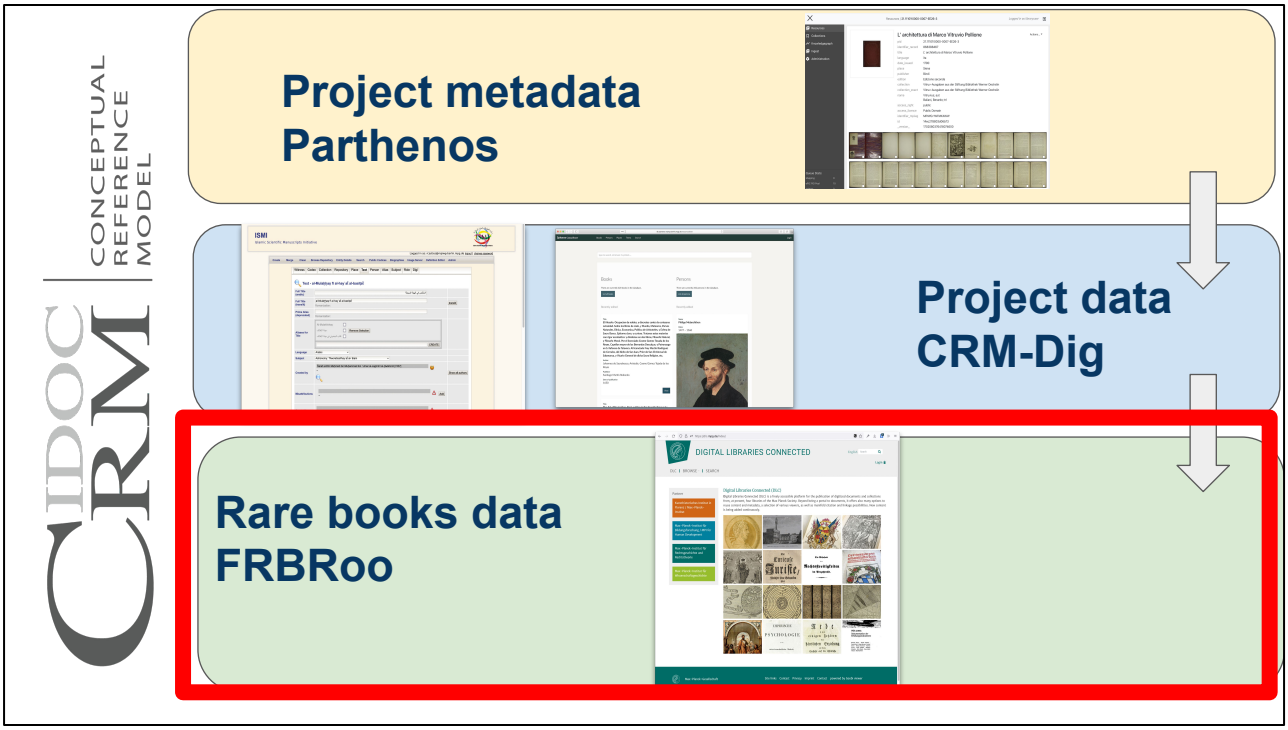
Managing Research Data Management



It does so by allowing us to describe the projects and actors, as well as the datasets, software and services provide. In particular, we keep track of where source data and mapping instructions have been archived that were used to generate the datasets in the Knowledge Graph. This image here is our own implementation model of the Parthenos ontology based on CIDOC CRM.

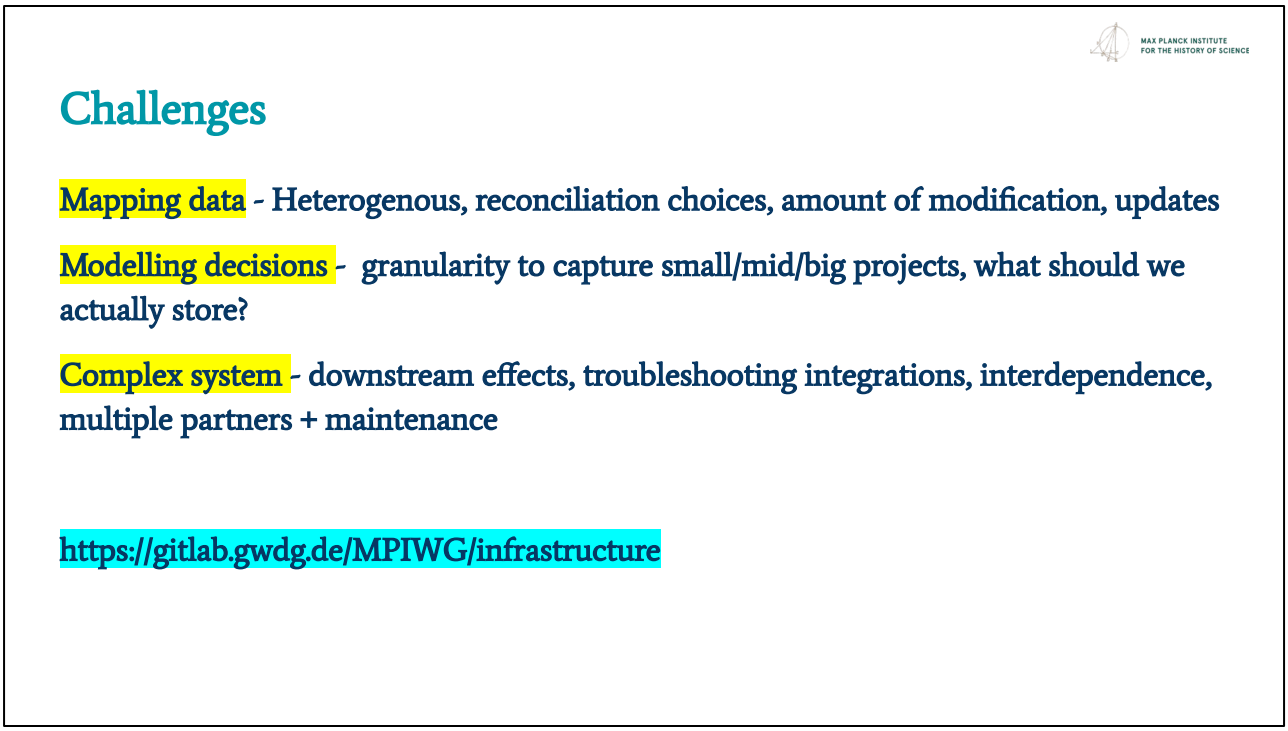


The project datasets themselves are stored in their native CIDOC CRM-based format or just the core elements such as Agents and Places are mapped using the X3ML tool.



The projects' research data is then linked back to the resources that are ingested from the library's rare book collection which is described with the FRBRoo extension of CIDOC-CRM.

And this is how finally the central knowledge graph becomes a rich environment that connects our digital humanities research data and the digitized documents from the library in one place.



We're still in the process of working through some big challenges, such as how much effort we want to put into reconciling heterogeneous data, and how much contextual data needs to be provided to make sense for us and our users.

We still need to think through how much of each project's research data and what aspects of it should be stored in our infrastructure, and if it is stored what is the process to normalize or align it to other data.

On top of the data and data modelling challenges we are faced with challenges in the complexity of maintaining the complex system and services on a technical level and organizational level, where the responsibility of different parts of the system are maintained by different groups and contractors. There are both downstream and upstream effects if there are issues in one system, which can be difficult to troubleshoot and often involve the cooperation of multiple partners to diagnose the issue.

Currently we're at the stage where we have a working proof of concept, but are continuing to test and gather usability feedback until the end of the year,

and we are looking forward to the next phase of development.

Thank you for listening!