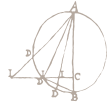


EASL, Oslo, 4-6 September 2019

RISE and SHINE: A distributed approach for text- centric research infrastructure

Sean Wang, Shih-Pei Chen

Max Planck Institute for the History of Science (MPIWG)



Current ecosystem for resource-centric research

Primary and **secondary** sources

Analog or **digital** (databases, e-books...)

External (commercial, freely available...)
or **internal** (produced by libraries themselves)
to the libraries





The trend in digital scholarship

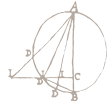
Scholars increasingly use **computational (and other more-than-read) methods** to analyze resources.

This trend may be challenging for libraries to support...

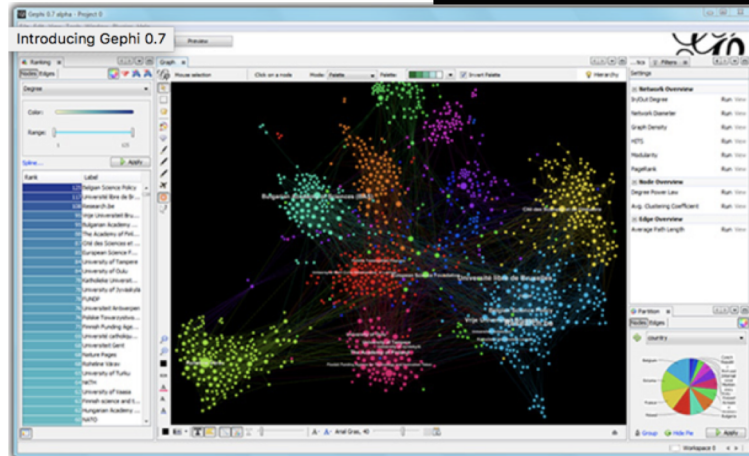
Technically, there are too many methods and tools, in different places and with different status, for libraries to provide in-house;

Legally, such usage often requires full access to large amount of computer-readable materials (e.g., “text-mining licenses” for commercial resources).

Existing research tools for general DH purposes



Network analysis



Mapping

GeoTemCo EXCELLENCE CLUSTER TOPOI PLATIN Place and Time Navigator

MAK-PLANCK-INSTITUT FÜR WISSENSCHAFTSGESCHICHTE DARIAH-DE

Load Data: KML File URL [] Load KML

Load Overlay: KML File URL [] Load KML

Data History: Heinrich Heine Save to Local Storage, Franz Kafka Save to Local Storage, Johann Wolfgang von Goethe Save to Local Storage

Background map: Open Street Map

Map selector tools: Filter, Publication Place

2,052 results

Heinrich Heine - place: remove

Franz Kafka - place: remove

Hamburg, Berlin, Leipzig, München, Amsterdam, Frankfurt am Main, Basel, Leipzig, München, New York, Berlin

Time start: 1765, Time unit: continuous, Scaling: normal, Animation: [] Dated Objects: 2052 results

Text analysis



N-gram, Regex, Replace, Similarity, Vectors, Diff, Transform

Word cloud, Chart, Save/Load, Help

Fetch text by URN: ctp:analects/li-ren Fetch Title: 厘仁 Qoen on ctext.org

子白: 「厘仁為美，擇不處仁，焉得知？」

子白: 「不仁者不可以久處約，不可以長處樂。仁者安仁，知者利仁。」

子白: 「唯仁者能好人，能惡人。」

子白: 「苟志於仁矣，無惡也。」

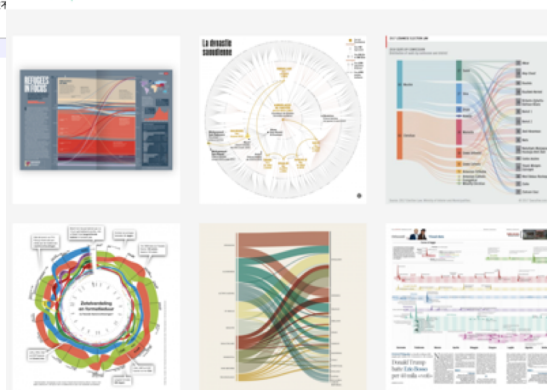
子白: 「富與貴是人之所欲也，不以其道得之，不處也；貧與賤是人之所惡也，不以其道得之，不處也。君子居則貴，而用則下，無所擇也。」

子白: 「我未見好仁者，惡不仁者，好仁者，無以尚之；惡不仁者，其為仁矣，不使後。」

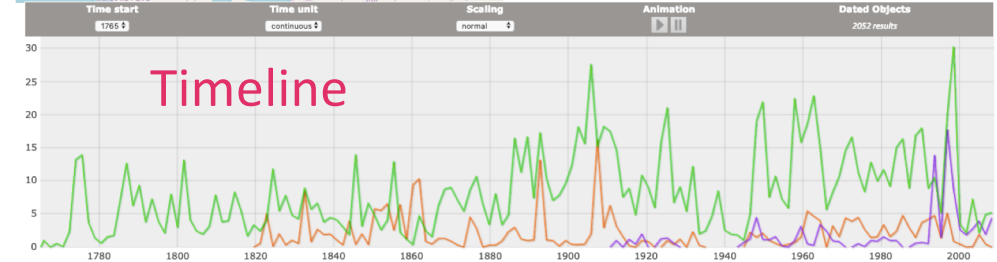
子白: 「人之過也，各於其黨。觀過，斯知仁矣。」

Save/add another text

RAWGraphs About Blog Learning Gallery



Timeline



Visual analytics

Text markup

Recogito | AN INITIATIVE OF Pelagios commons

MARKUS

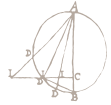
With MARKUS you can upload a file in classical Chinese (and perhaps in the future of personal names, place names, temporal references, and bureaucratic office) upload your own list of key terms for automated tagging. You can then read a document reference works at the same time, or compare passages in which the same names extract the information you have tagged and use it for further analysis in our visualization.

Step 1: Upload a txt (UTF-8) or saved MARKUS file or Paste your txt here

Semantic Annotation without the pointy brackets

Work on texts and images. Identify and mark named entities. Use your data in other tools or connect to other data on the Web. Without the need to learn code.

Learn More



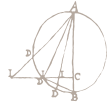
How could libraries better support digital scholarship?

Possible scenario #1: A non-reproducible solution to working with (commercial) resources

Publishers sometimes work with scholars on an *ad hoc* basis to provide **downloadable ‘raw data’**. Some cultural institutions also provide data in this manner.

Challenging to replicate on a large scale, and makes reproducibility more difficult.

Risks losing track of content’s lifecycle (e.g., versions) and further (re-)usage.



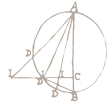
How could libraries better support digital scholarship?

Possible scenario #2: Big silos

Some large research infrastructures provide encapsulated environments for scholars to use computational methods on open-access or commercial resources (e.g., HathiTrust, TextGrid)

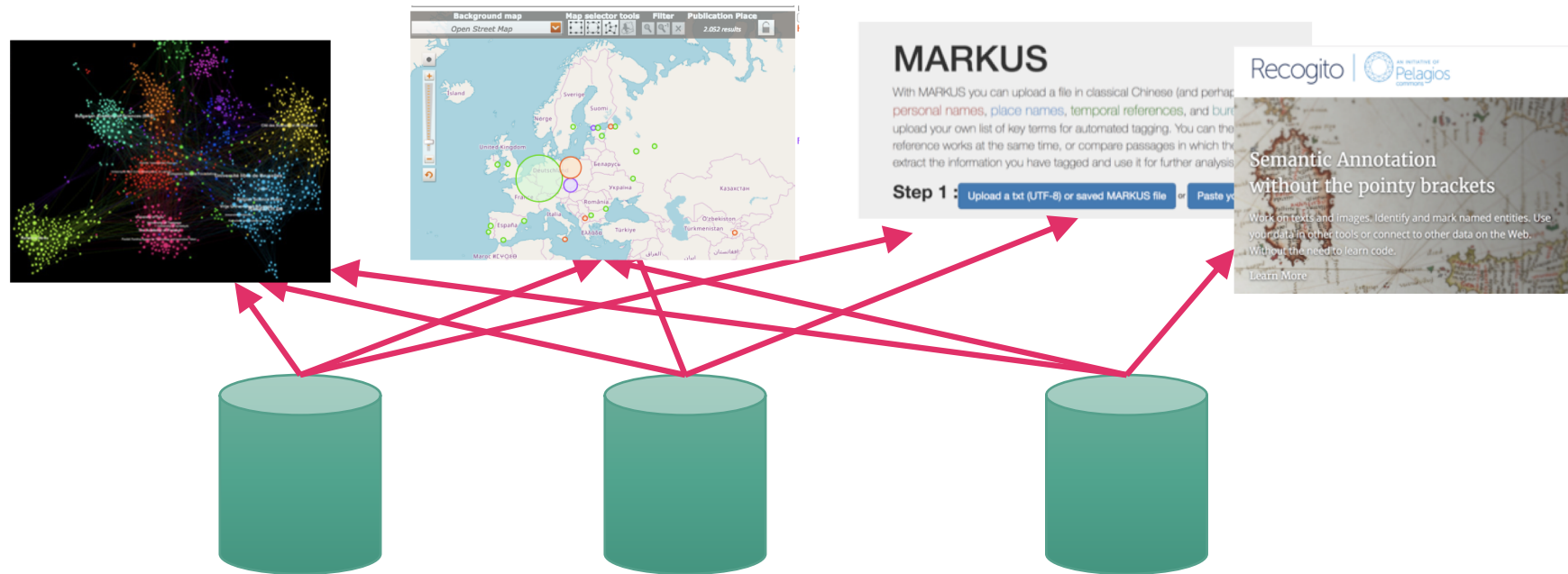
- ⇒ **Open platforms** cannot work with commercial resources.
- ⇒ **Closed platforms** (such as encapsulated environments) suffer from limited tool coverage and challenging licensing negotiations.

For **scholars**, such silos are limiting in terms of both available resources and tools.



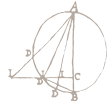
Our proposal: a distributed network of resources and tools

Linking contents with research tools in place, creating a ***distributed network*** of resources and tools.



Scholars can ***choose resources from any databases*** and ***send them to any tools of choice*** to analyze, if they have proper rights.

Our proposal: RISE and SHINE



<https://rise.mpiwg-berlin.mpg.de>

A common **API** (a language that everyone talks) for transferring texts between databases and tools => **SHINE**

A secure and trust-worthy **guard in between every transfer** to make sure a text is transferred to an authenticated user from an authorized institute => **RISE**

Resource Providers

RISE authorization

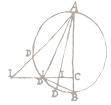
Research Tools

Resource Providers

RISE Authorization

Research Tools





The RISE infrastructure

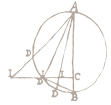
A suite of software packages covering the needs of different stakeholders

A *middleware* that catalogs all linked resources, and authenticates and authorizes text transfers

Management interfaces for libraries / resource providers, research institutes, and tool developers to set privileges

A suite of *JavaScript libraries* to allow easy integration with the SHINE API **for software developers**

A *Resource Provider* software package that allows **resource providers** (database owners, archival institutions, or **even scholars**) to share resources in a protected, SHINE-compatible way



Challenges: how to make the RISE middleware work for libraries?

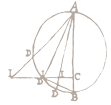
How to work with libraries' **internal database management systems** automatically?

- To avoid duplication of efforts!

How to make **authentication and authorization** seamless?

- Options beyond Shibboleth and RISE's own user registry?

How to technically represent **the entire spectrum of licensing rights**, from fully open to completely protected?



Call for collaboration

This should be a **network built by the community**

- Work together to define this network
- Call for collaboration with libraries to test this concept!

RISE will allow **libraries** to offer digital scholarship with both licensed and open resources!

Check our website for detailed documentations, API, & available toolkits

- <https://rise.mpiwg-berlin.mpg.de>