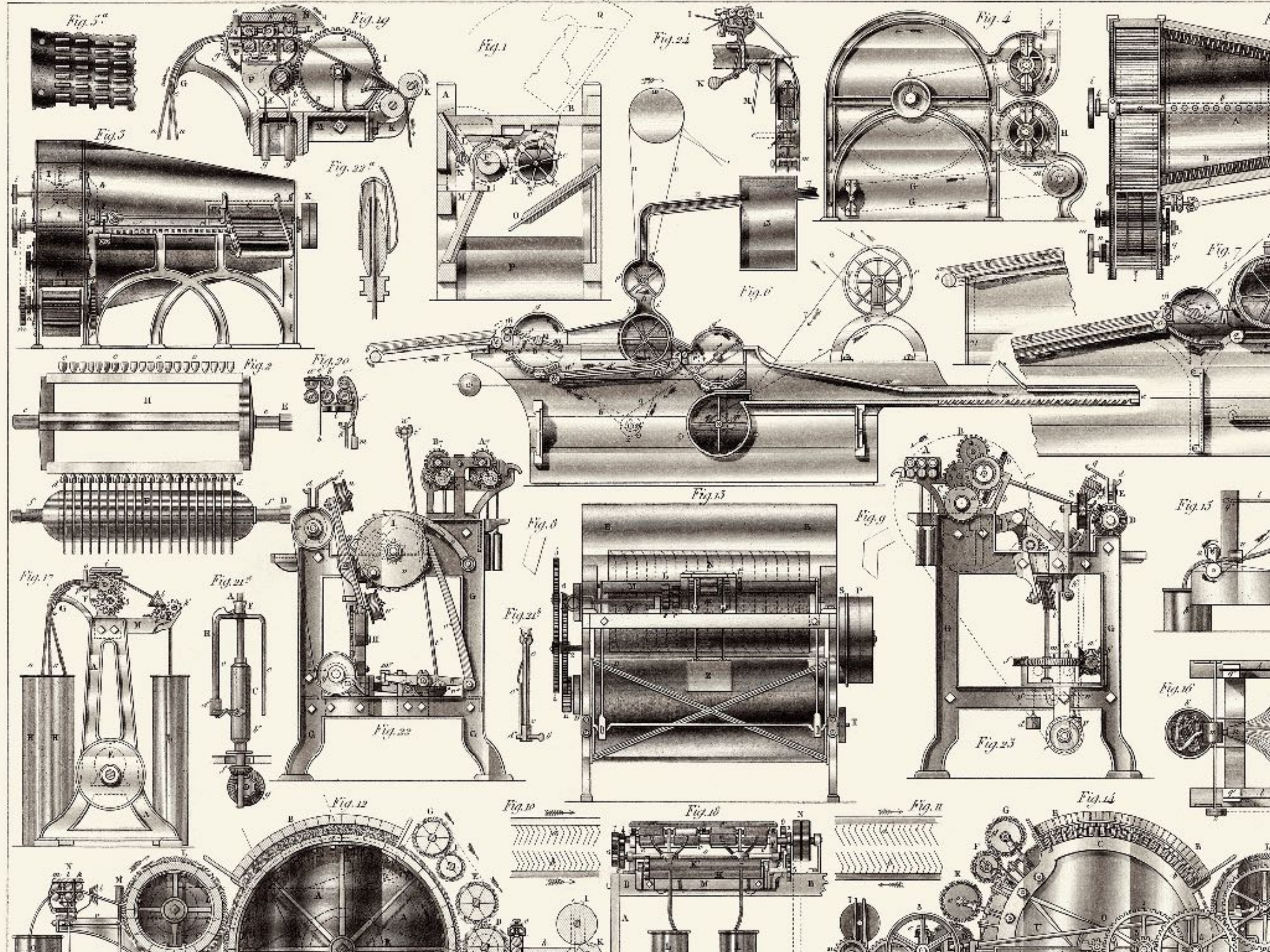


Modeling the Past—

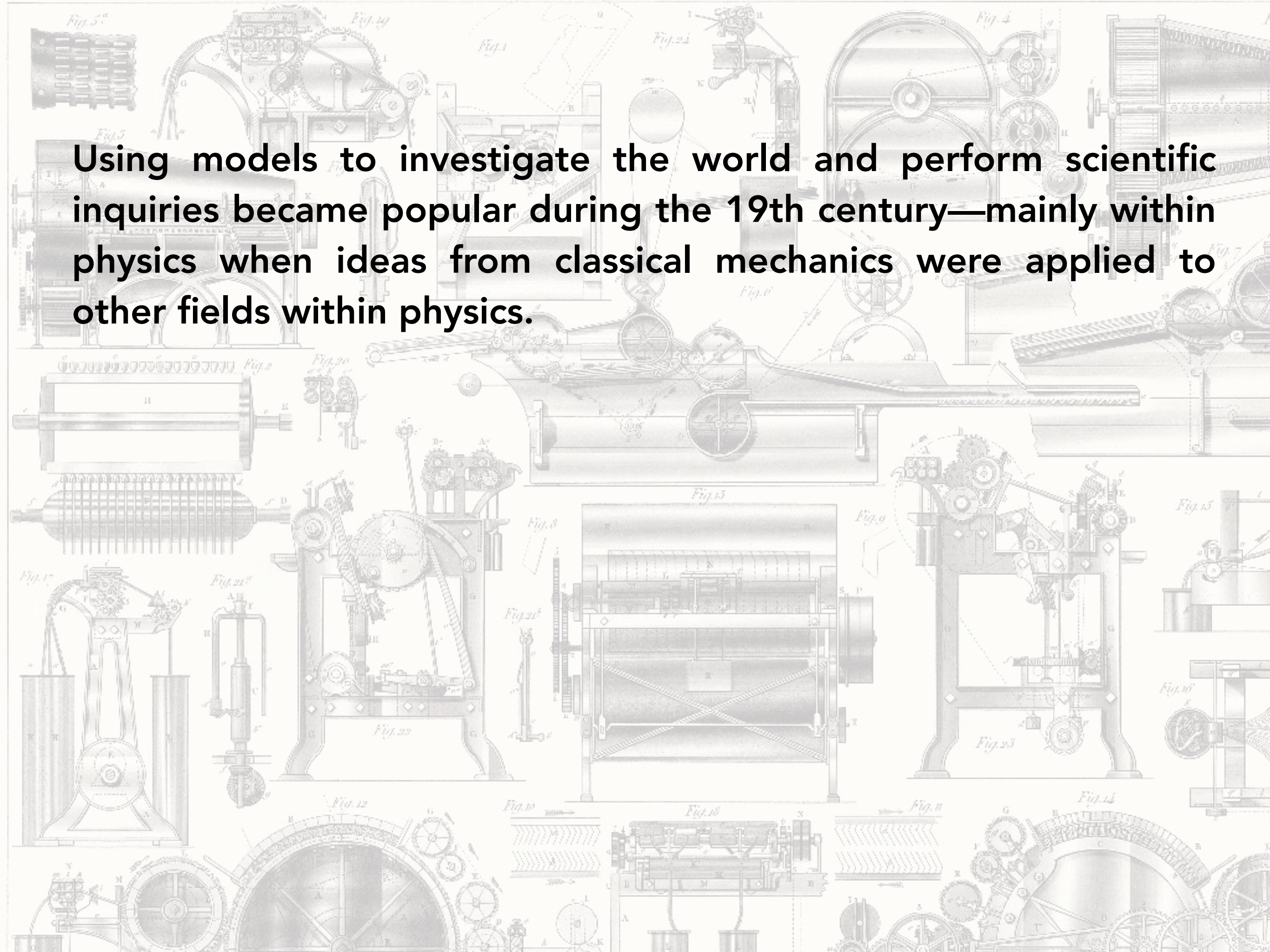
# Between the History of Technology & the Digital Humanities

Prof. Pelle Snickars  
Department of Culture and Media Studies / Humlab  
Umeå University









Using models to investigate the world and perform scientific inquiries became popular during the 19th century—mainly within physics when ideas from classical mechanics were applied to other fields within physics.



Essentially, the idea—or concept—of a **model** has three meanings:



Essentially, the idea—or concept—of a **model** has three meanings:

First of all, the Latin word *modellus*—stemming from *modulus*—refers to a **small measuring device**.



Essentially, the idea—or concept—of a **model** has three meanings:

First of all, the Latin word *modellus*—stemming from *modulus*—refers to a **small measuring device**.

Secondly, the word “model” also has an everyday meaning—how something should look like or **how some procedure ought to be carried out**.



Essentially, the idea—or concept—of a **model** has three meanings:

First of all, the Latin word *modellus*—stemming from *modulus*—refers to a **small measuring device**.

Secondly, the word “model” also has an everyday meaning—how something should look like or **how some procedure ought to be carried out**.

Finally, more explicit **scientific models function as simplifications**.



Philip Gerlee  
Torbjörn Lundh

# Scientific Models

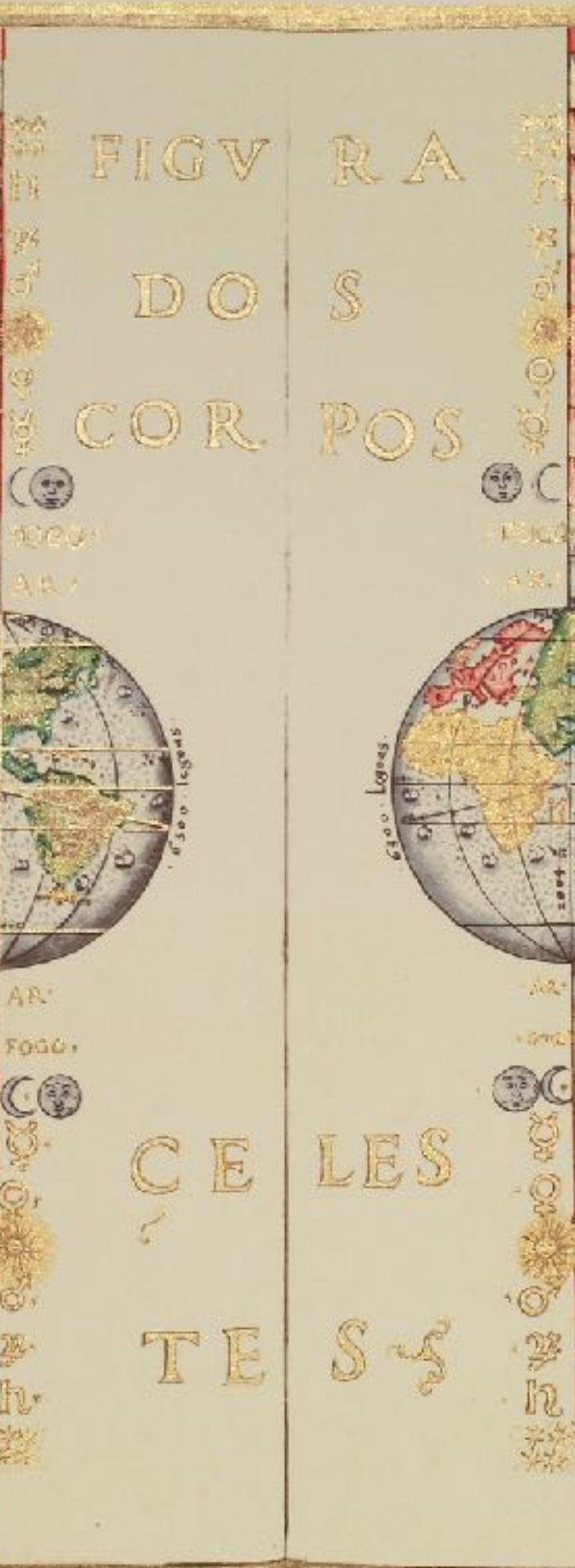
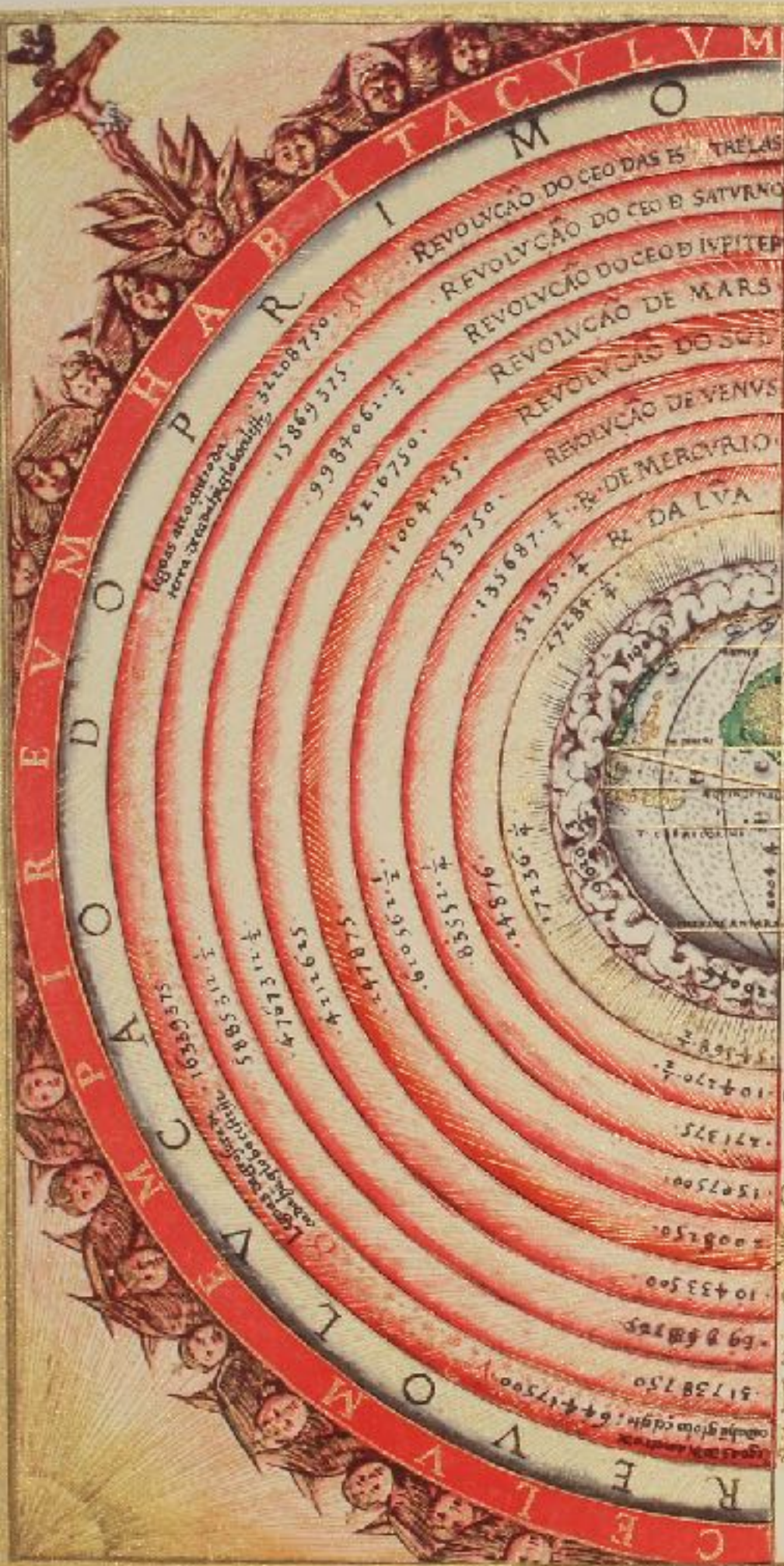
Red Atoms, White Lies  
and Black Boxes in a Yellow Book

 Springer

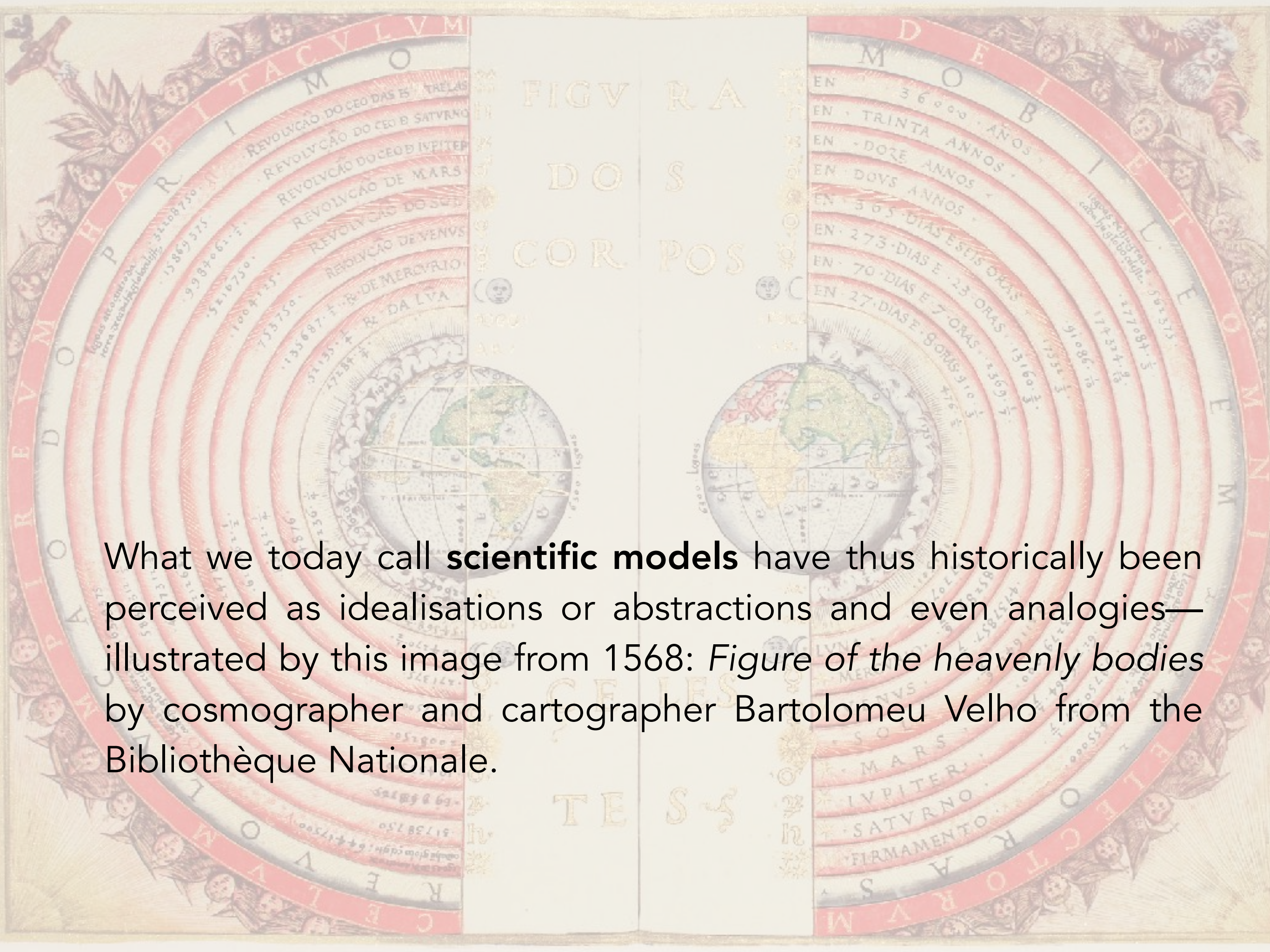


Gerlee and Lund argues that the word model also applies—at least to some degree—to world views that “**existed before the scientific revolution of the 18th century**, such as the geocentric model of the solar system. An important difference between this use and the modern concept of a model is that the scientists of antiquity and the Middle Ages didn’t consider their models as simplifications or idealisation of a more complex world, but rather as **direct representations of reality**. In this sense the pre-modern concept of a model lies closer to models as ideals, and not models as simplifications.”





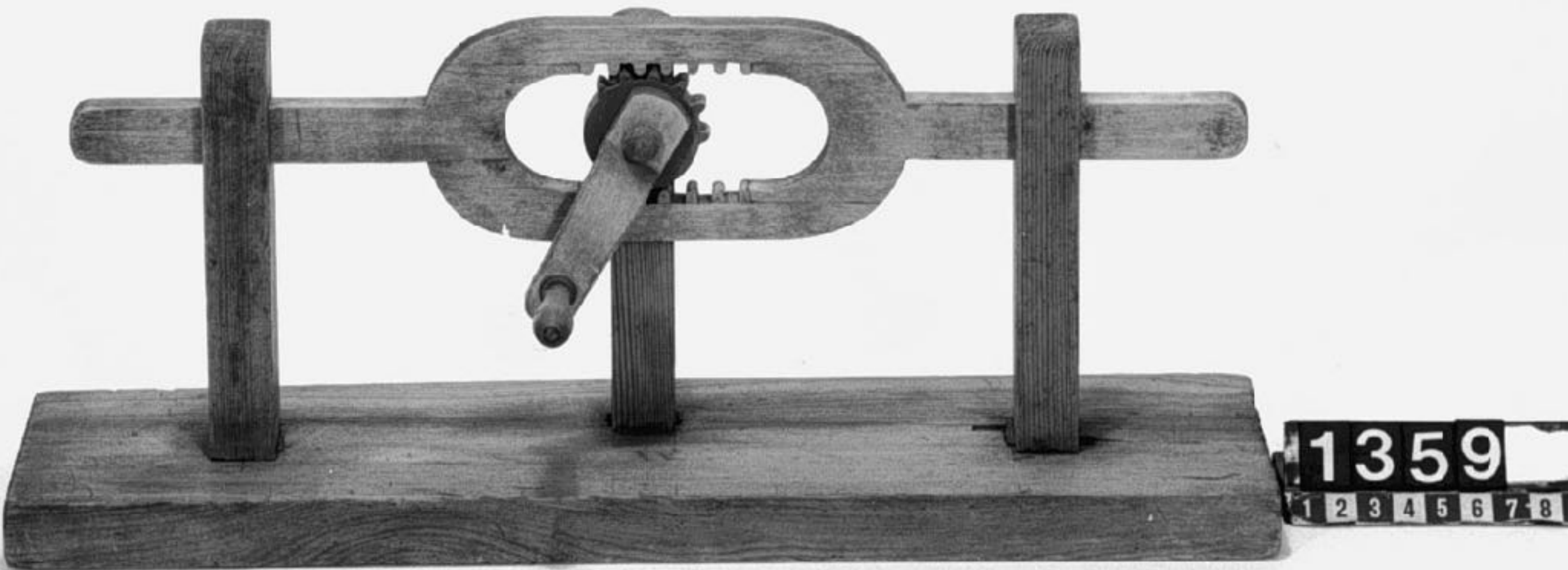




What we today call **scientific models** have thus historically been perceived as idealisations or abstractions and even analogies—illustrated by this image from 1568: *Figure of the heavenly bodies* by cosmographer and cartographer Bartolomeu Velho from the Bibliothèque Nationale.

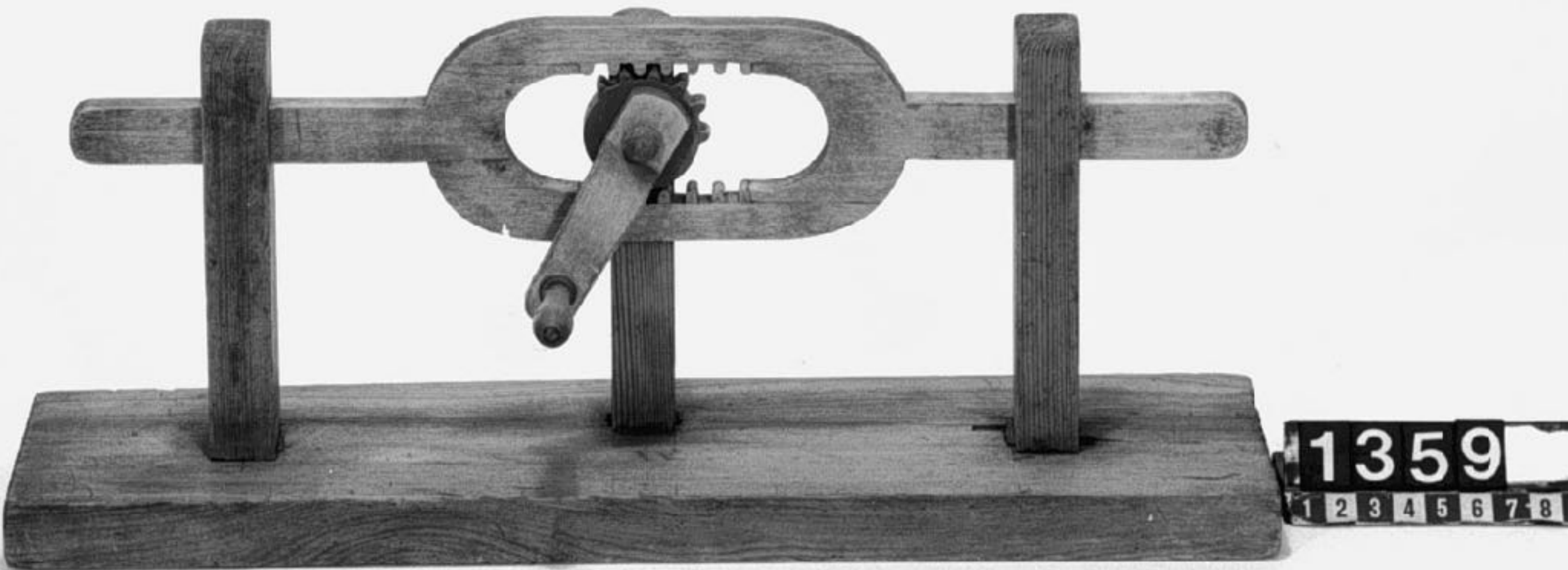


However, during the 18th century these ideas were gradually transformed—and the notion of models also began to refer to **small-scale physical manifestations.**





Importantly, **the actual word "model"** has had its current meaning only since the beginning of the 20th century—before the word was used exclusively to denote **actual physical models**.





# Models as Illustrations





Models as Illustrations

Illustrations as Models

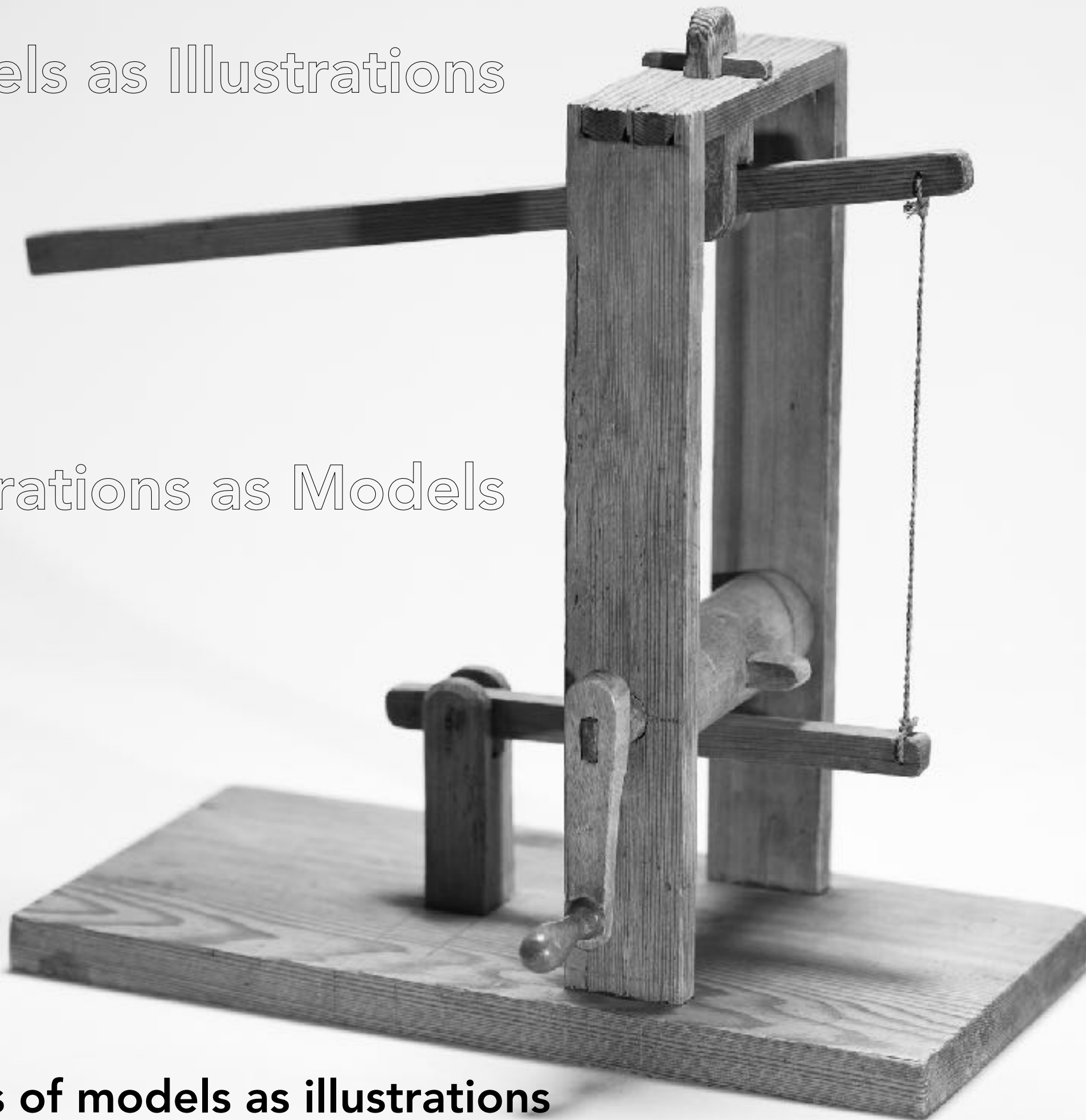




Models as Illustrations

Illustrations as Models

**Images of models as illustrations**



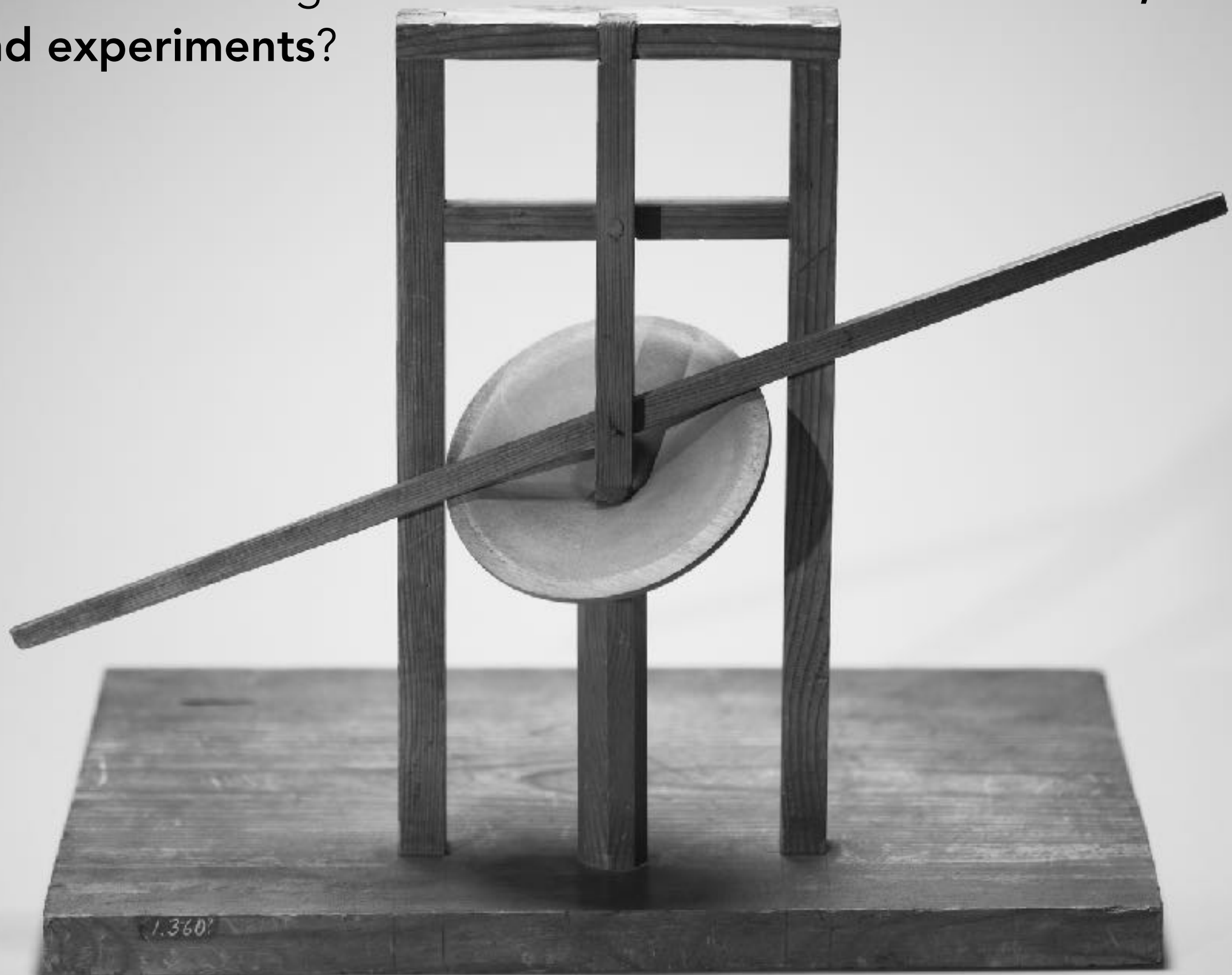


Yet, if scientific models function as simplifications—does such a perspective **apply for history** as well?





Can the **history of models or modelling practices**, for example, tell us something about **the relation between models, theory and experiments?**





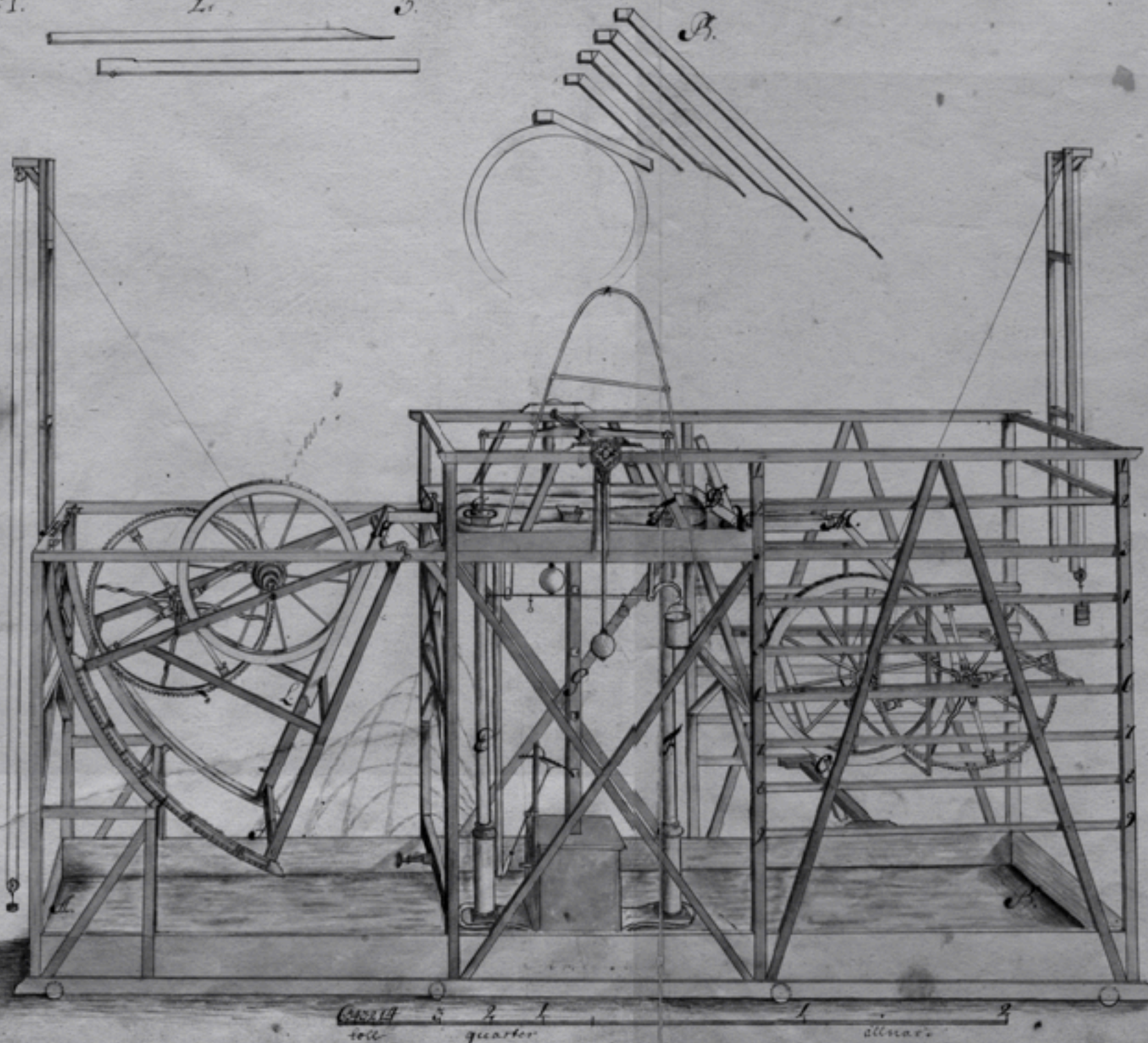
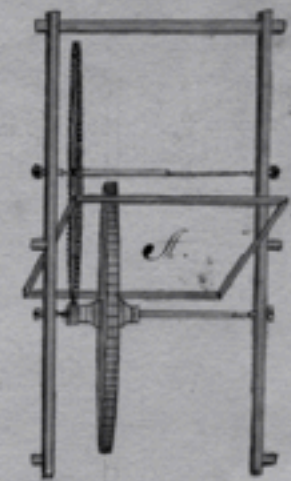
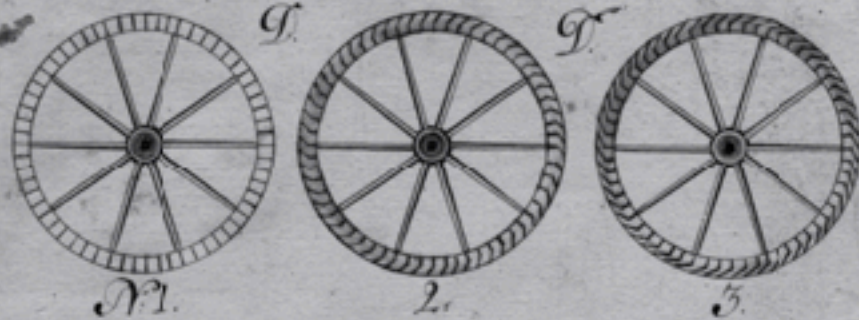
Modeling the Past—

**Between the History of Technology & the Digital Humanities**



Den första Machinen om Ballnings Kraft i tryckfulla fall.

N: 1.



Christopher Polhem's hydro-dynamic "experimental machine" for water pressure measurements (1705).



# 1. Introduction

2. From Archival to Data Driven Humanistic Research

3. About the Research Project "Digital Models"

4. Textual Models of the Past

5. Visual Models of the Past

6. Conclusion

All slides in the form of a PDF can be found at <http://pellesnickars.se/>



The background is a historical manuscript page, likely from a technical or scientific work. At the top, there is a title in Swedish: "Den första Machinen om Wallnuths Kraft i ägghällige fall." (The first machine about the power of the nut in the egg-shaped fall). The page is numbered "N: 1." in the top right corner. The main content consists of several technical drawings. On the left, there are three wheels labeled 1, 2, and 3. In the center, there are two cross-shaped mechanical components labeled 4 and 5. To the right of these, there are three small mechanical parts labeled 6, 7, and 8. On the far right, there is a large, complex mechanical structure, possibly a pump or a mill, labeled 9. The drawings are detailed and show various mechanical components like gears, levers, and frames. The text is written in a cursive script, typical of the 18th or 19th century.

1. Introduction

**2. From Archival to Data Driven Humanistic Research**

3. About the Research Project "Digital Models"

4. Textual Models of the Past

5. Visual Models of the Past

6. Conclusion

All slides in the form of a PDF can be found at <http://pellesnickars.se/>





1. Introduction

2. From Archival to Data Driven Humanistic Research

**3. About the Research Project "Digital Models"**

4. Textual Models of the Past

5. Visual Models of the Past

6. Conclusion

All slides in the form of a PDF can be found at <http://pellesnickars.se/>





1. Introduction

2. From Archival to Data Driven Humanistic Research

3. About the Research Project "Digital Models"

**4. Textual Models of the Past**

5. Visual Models of the Past

6. Conclusion

All slides in the form of a PDF can be found at <http://pellesnickars.se/>



The background is a historical manuscript page, likely from a technical or scientific work. At the top, there is a title in Swedish: "Den första Machinen om Wallnuths Kraft i ägghällige fall." (The first machine about the power of the nut in the egg-shaped fall). The page is numbered "N: 1." in the top right corner. The main content consists of several technical drawings. On the left, there are three wheels labeled 1, 2, and 3. In the center, there are several smaller mechanical components, including a cross-shaped part labeled B. On the right, there is a large, complex machine with a frame and various mechanical parts. The machine is shown in a perspective view, with a large wheel on the left and a smaller wheel on the right. The machine is supported by a frame with diagonal bracing. At the bottom, there is a scale bar with markings for "toil", "quarter", and "all over".

1. Introduction

2. From Archival to Data Driven Humanistic Research

3. About the Research Project "Digital Models"

4. Textual Models of the Past

**5. Visual Models of the Past**

6. Conclusion

All slides in the form of a PDF can be found at <http://pellesnickars.se/>



The background is a historical manuscript page, likely from a 17th-century technical treatise. At the top, there is a title in Swedish: "Den första Machinen om Wallnuths Kraft i ägghällige fall." (The first machine about the strength of the walnut in the egg-shaped fall). The page is numbered "N: 1." in the top right corner. The manuscript contains several detailed technical drawings. At the top left, there are three wheels labeled 1, 2, and 3. To their right are various mechanical components, including a cross-shaped part labeled B, a conical part labeled F, a cylindrical part labeled G, and a long rod labeled H. On the far right, there is a drawing of a frame structure. The bottom half of the page is dominated by a large, complex drawing of a machine, which appears to be a type of press or mill, with a large wheel and a frame. The machine is labeled with letters A, B, C, D, E, F, G, H, I, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z. The machine is shown in a perspective view, with a large wheel on the left and a frame on the right. The drawing is very detailed, showing the internal mechanisms and the structure of the machine. The text is written in a cursive script, typical of the 17th century. The overall tone of the image is historical and technical.

1. Introduction

2. From Archival to Data Driven Humanistic Research

3. About the Research Project "Digital Models"

4. Textual Models of the Past

5. Visual Models of the Past

**6. Conclusion**


All slides in the form of a PDF can be found at <http://pellesnickars.se/>






# Modeling the Past



2017-10-10
Modeling the Past. Predictive Approaches to Climate, Environment, Population and cultural Dynamics in Prehistory | Topoi

EXCELLENCE  
CLUSTER

TOPOI

THE FORMATION AND TRANSFORMATION OF SPACE AND  
KNOWLEDGE IN ANCIENT CIVILIZATIONS

FOLLOW US

INTRANET

SEARCH


HOME | RESEARCH | KNOWLEDGE TRANSFER | PEOPLE | PARTNERS | CAREER

ABOUT US | CONTACT | SERVICE | NEWS | CALENDAR | BERLINER ANTIKE-KOLLEG | EDITION | TOPOI

Workshop

# MODELING THE PAST

## PREDICTIVE APPROACHES TO CLIMATE, ENVIRONMENT, POPULATION AND CULTURAL DYNAMICS IN PREHISTORY


<p><b>Speakers:</b></p> <p>Carsten Lemmen (Helmholtz Zentrum Geestacht)</p> <p>Wolfgang Schwanghard (University of Potsdam)</p> <p>Brian Beckers (FU Berlin)</p> <p>Ulrich Cubasch (FU Berlin)</p> <p>Janina Körper (FU Berlin)</p> <p>Emmanuele Russo (FU Berlin)</p> <p>B. Fallah (FU Berlin)</p> <p>Jed Kaplan (University of Geneva)</p> <p>Jutta Lechterbeck (Landesdenkmalamt Baden-Württemberg, Hemmenhofen)</p> <p>Tim Evans (Imperial College London)</p> <p>Andreas Zimmermann (University of Cologne)</p> <p>Mehdi Saqalli (Université Toulouse le Mirail)</p>	<p><b>WHEN</b></p> <p>21.11.2013 - 22.11.2013</p> <p><b>LOCATION</b></p> <p>Topoi Building Dahlem Hitdorfsstraße 18 14195 Berlin Deutschland <a href="#">→ google maps ↗</a></p> <p><b>ORGANISER</b></p> <p>Wolfram Schier <a href="mailto:wolfram.schier@topoi.org">wolfram.schier@topoi.org</a> +49 30 838-55873</p> <p>Brigitta Schütt <a href="mailto:brigitta.schuett@fu-berlin.de">brigitta.schuett@fu-berlin.de</a> +49 30 838-70479</p> <p><b>DOWNLOADS</b></p> <p><a href="#">Program (55.57 Kb)</a></p> <p><b>RESEARCH AREA</b></p> <p>(A) Spatial Environment</p>
---	--

Contact | Impressum

This site uses Piwik<sup>®</sup> to statistically evaluate visitor behaviour.  
☒ You are currently opted in. [Click here to opt out.](#)

<https://www.topoi.org/event/22272/>
1/1


Have library access? [Login Through Your Library](#)



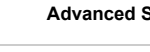
[Login to My Account](#) | [Register](#)

[Advanced Search](#)
[Browse](#)

[About](#)
[Support](#)
[Login](#) | [Register](#)



The MIT Press



**JOURNAL ARTICLE**

**Modeling the past: The Specification of Functional Form**

Ivy Broder and Allan J. Lichtman

*The Journal of Interdisciplinary History*

Vol. 13, No. 3 (Winter, 1983), pp. 489-502

Published by: [The MIT Press](#)

DOI: 10.2307/202947

Stable URL: <http://www.jstor.org/stable/202947>

Page Count: 14

**Topics:** [Regression analysis](#), [Mathematical independent variables](#), [Mathematical dependent variables](#), [Modeling](#), [Linear regression](#), [Coefficients](#), [Voting](#), [Linear models](#), [Regression coefficients](#), [Voting behavior](#)

Were these topics helpful? [See something inaccurate? Let us know!](#)

Viewing page [489] of pages 489-502

*Journal of Interdisciplinary History*, XIII:3 (Winter 1983), 489-502.

Ivy Broder and Allan J. Lichtman

---

**Modeling the Past: The Specification of Functional Form** Increasingly historians and other social scientists are relying on regression analysis for assessing the influence of variables on a particular behavior of interest. Unlike correlation coefficients, which only indicate the strength and direction of the association between variables, regression measures disclose how the values of one variable (termed the dependent





## Modeling the Past: Digital Technologies and Excavations in Polis, Cyprus

Euro-Mediterranean Conference

EuroMed 2012: Progress in Cultural Heritage Preservation pp 414-422 | Cite as

- Joanna S. Smith (1)
- Szymon M. Rusinkiewicz (2)

1. Department of Art and Archaeology, Princeton University, USA
2. Department of Computer Science, Princeton University, USA

Conference paper

- [6 Readers](#)
- [3.1k Downloads](#)

Part of the [Lecture Notes in Computer Science](#) book series (LNCS, volume 7616)

### Abstract

This research and educational project aimed to create virtual 3-D walkthroughs of four principal buildings from the Princeton University excavations at Polis Chrysochous, Cyprus. The structures date from the Cypro-Archaic period beginning in the 7th century bce to the Late Antique period of the 7th century ce. The project was conceived together with a special exhibition, a long-term exhibition in Cyprus, and a presentation on the web. In a joint Computer Science and Art and Archaeology seminar in the spring of 2012, students created reconstructions and populated them with 3-D scanned objects. The challenge was to find appropriate visual metaphors for conveying uncertainty and change in these 3-D visualizations as well as to create a computer-animated movie focused on the buildings, their spatial relationships, and possible reconstructions consistent with the excavations.

### Keywords

3-D digital modeling 3-D scanning archaeology Arsinoe Cyprus excavation exhibition Marion museum Polis Chrysochous public students

### Preview

The Center for the Humanities / Programming /

## Modeling the Past

EVENT

Wed, Oct 24, 2012, 06:30 PM – 06:30 PM



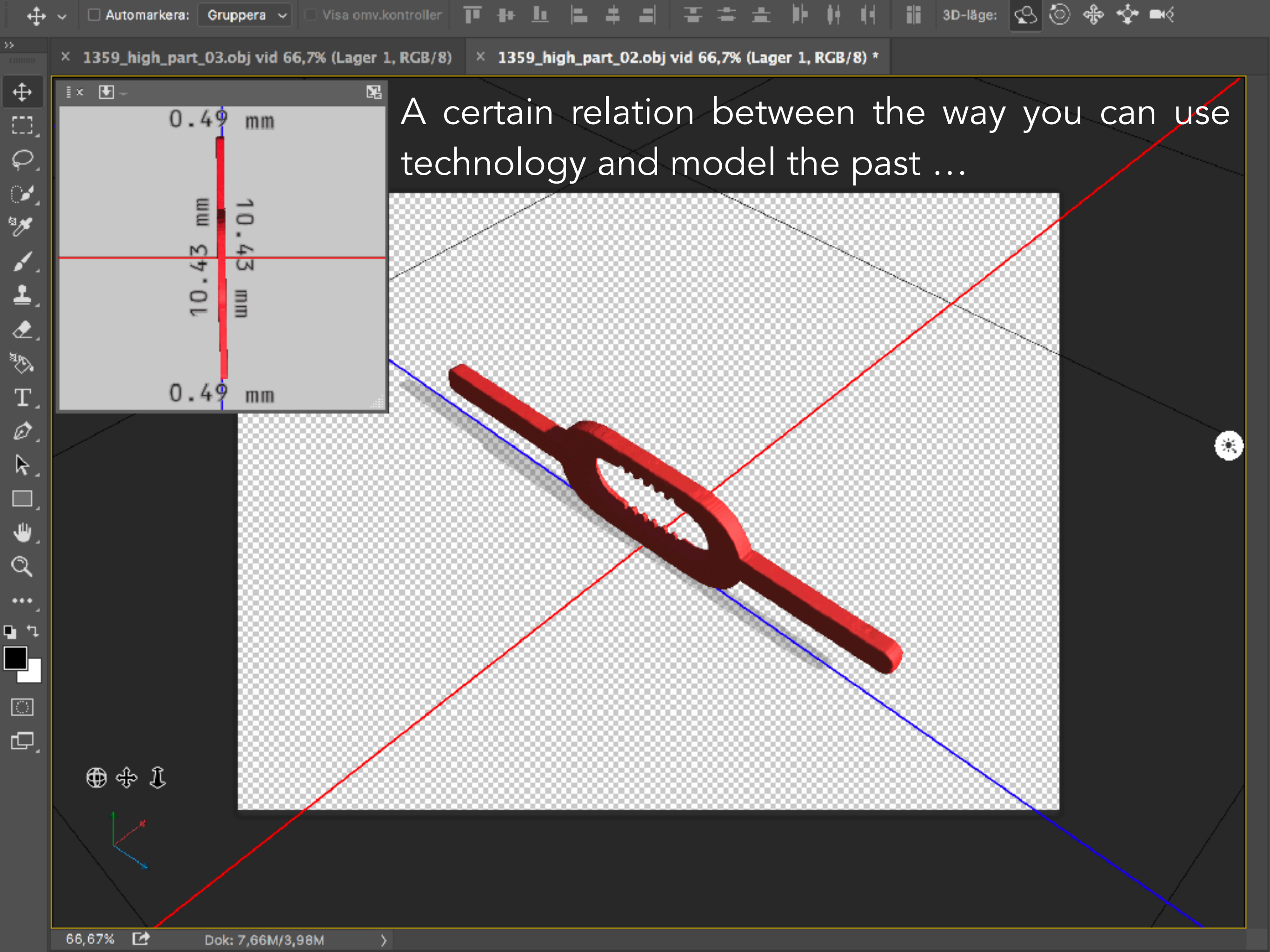
Source:WikiMedia Commons

### About the event

With generous funding from an anonymous donor and the National Science Foundation, and in close cooperation with the *Soprintendenza per i Beni Archeologici del Lazio*, an international team has been creating a restoration model of Hadrian's Villa, a World Heritage Site, and the best-preserved imperial villa in the hinterland of Rome. The model includes terrain, gardens, water features, sculpture, buildings, furnishings, and avatars representing members of the imperial court. The IDIA Lab at Ball State University, a partner in the project, has taken the 3D model and ported it to the game engine, Unity3D, so that it is possible to explore the reconstructed villa interactively over the Internet. This talk will present the project, its history, goals, current state, and future prospects.

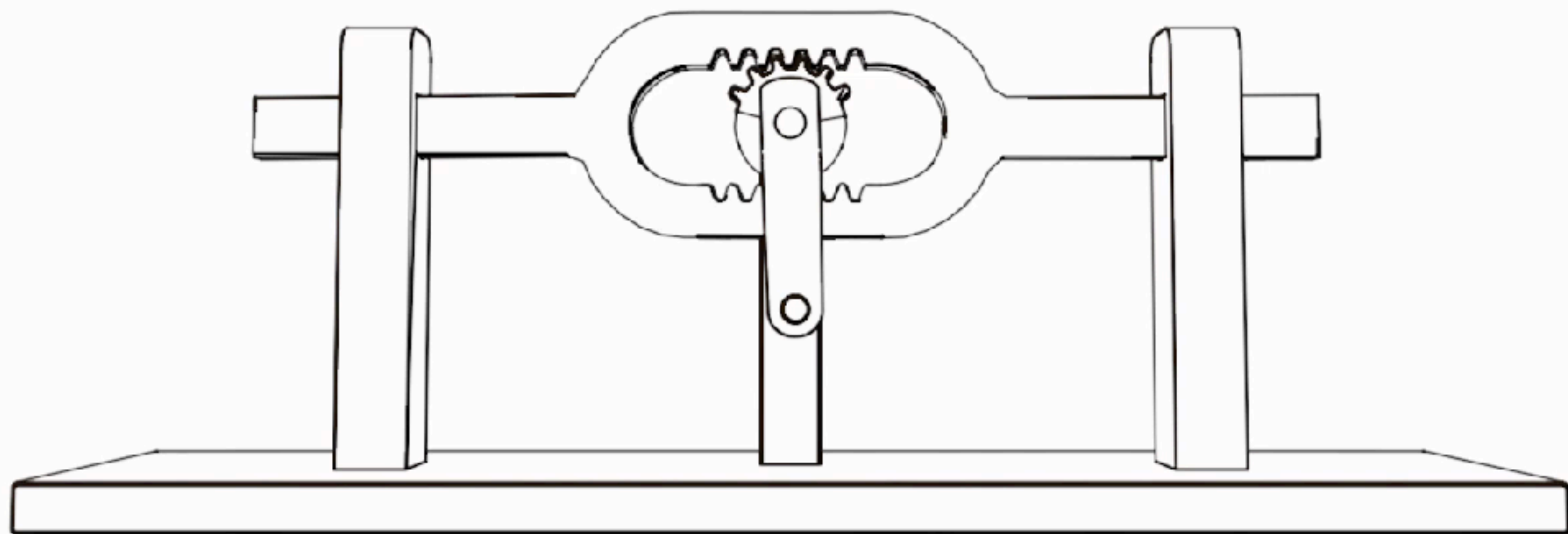
### Media



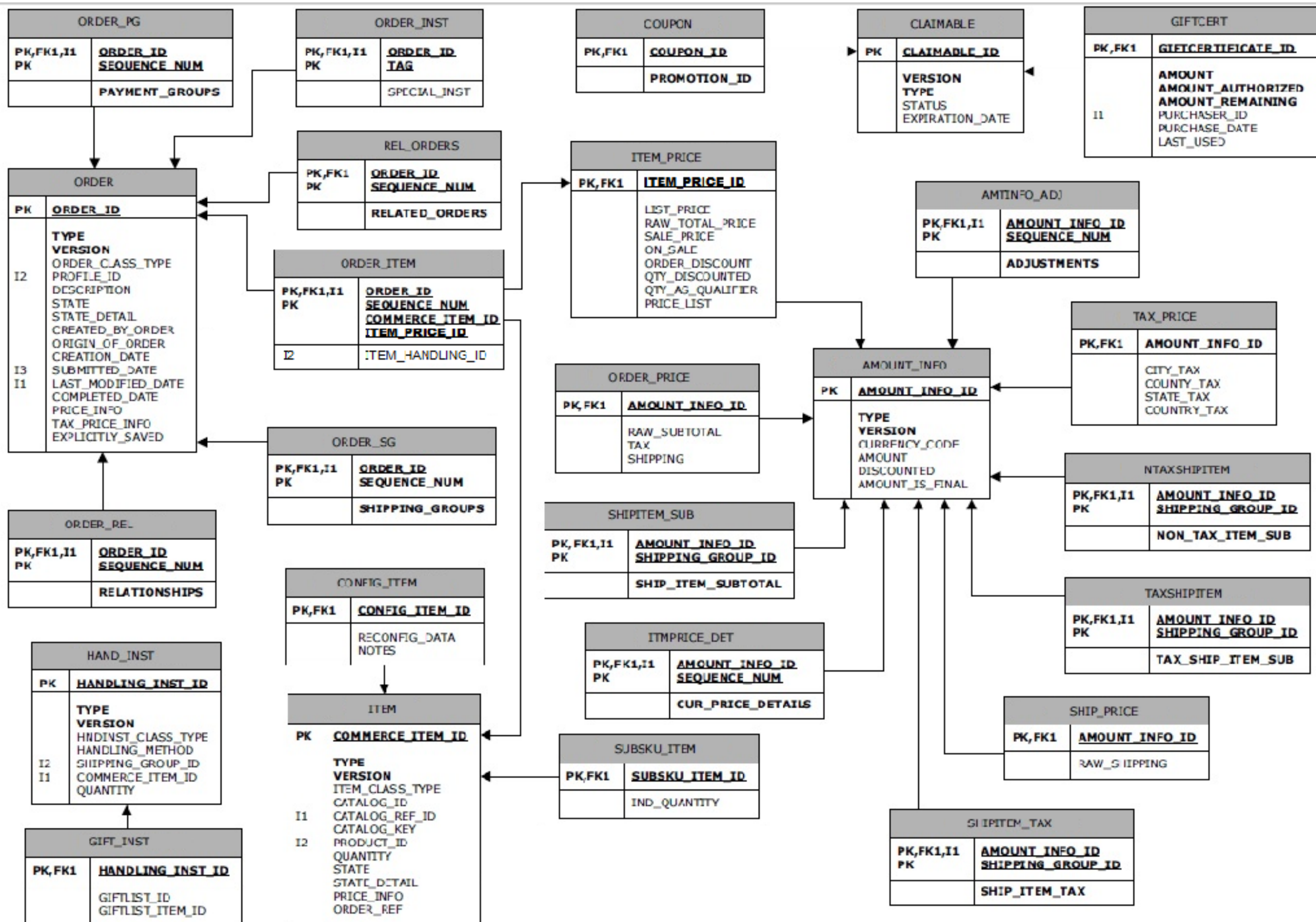


A certain relation between the way you can use technology and model the past ...









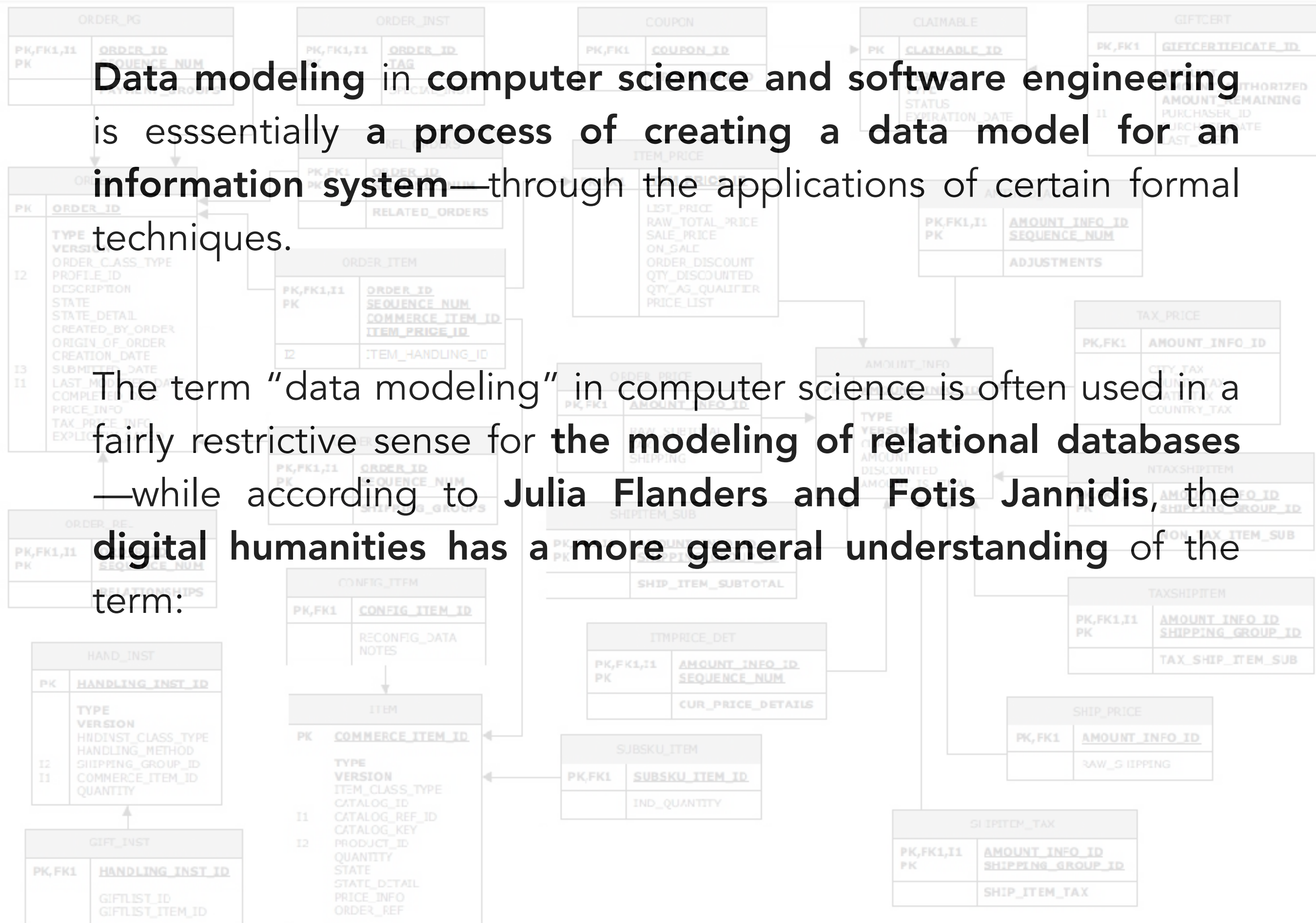






Data modeling in computer science and software engineering is essentially a process of creating a data model for an information system—through the applications of certain formal techniques.

The term “data modeling” in computer science is often used in a fairly restrictive sense for the modeling of relational databases—while according to Julia Flanders and Fotis Jannidis, the digital humanities has a more general understanding of the term:





**"Data modeling** is the modeling of some segment of the world in such a way to make some aspects computable, referring to creating database schemas ... XML schemas [or] ontologies."

Flanders & Jannidis white paper, "**Knowledge Organization and Data Modeling in the Humanities**" (2015) – <https://hcommons.org/deposits/item/hc:12445/>



**Project**

---

Source

---

Issues

---



**word2vec**

---

Tool for computing continuous distributed representations of words.





Code | Archive

**Projects**

Search

About

**Project**

---

Source

---

Issues

---



**word2vec**

---

Tool for computing continuous distributed representations of words.



To observe strong regularities in the word vector space, it is needed to train the **models** on large data set, with sufficient vector dimensionality as shown in [1]. Using the **word2vec** tool, it is possible to train **models** on huge data sets (up to hundreds of billions of words).

... to train models ...

<https://code.google.com/archive/p/word2vec/>



So called “**word-embedding models**” basically computes words into **pure numbers**—representing **semantics as mathematical vectors**.

A word-embedding model embeds a word in a space that **represent** semantic and syntactic relationships between words.



So called “**word-embedding models**” basically computes words into **pure numbers**—representing **semantics as mathematical vectors**.

A word-embedding model embeds a word in a space that **represent** semantic and syntactic relationships between words.

---

## Efficient Estimation of Word Representations in Vector Space

---

**Tomas Mikolov**  
Google Inc., Mountain View, CA  
tmikolov@google.com

**Kai Chen**  
Google Inc., Mountain View, CA  
kaichen@google.com

**Greg Corrado**  
Google Inc., Mountain View, CA  
gcorrado@google.com

**Jeffrey Dean**  
Google Inc., Mountain View, CA  
jeff@google.com

### Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

### 1 Introduction

Many current NLP systems and techniques treat words as atomic units - there is no notion of similarity between words, as these are represented as indices in a vocabulary. This choice has several good reasons - simplicity, robustness and the observation that simple models trained on huge amounts of data outperform complex systems trained on less data. An example is the popular N-gram model used for statistical language modeling - today, it is possible to train N-grams on virtually all available data (trillions of words [3]).

However, the simple techniques are at their limits in many tasks. For example, the amount of relevant in-domain data for automatic speech recognition is limited - the performance is usually dominated by the size of high quality transcribed speech data (often just millions of words). In machine translation, the existing corpora for many languages contain only a few billions of words or less. Thus, there are situations where simple scaling up of the basic techniques will not result in any significant progress, and we have to focus on more advanced techniques.

With progress of machine learning techniques in recent years, it has become possible to train more complex models on much larger data set, and they typically outperform the simple models. Probably the most successful concept is to use distributed representations of words [10]. For example, neural network based language models significantly outperform N-gram models [1, 27, 17].

#### 1.1 Goals of the Paper

The main goal of this paper is to introduce techniques that can be used for learning high-quality word vectors from huge data sets with billions of words, and with millions of words in the vocabulary. As far as we know, none of the previously proposed architectures has been successfully trained on more



## word2vec

In these models **words are represented as vectors** along a few hundred artificial dimensions—that **do not correspond** to documents or other real textual contexts. It is **pure numbers—words become vectors**.

Some word-embedding models are therefore **prediction-based**.



## Word Vectors in the Eighteenth Century

Ryan James Heuser  
heuser@crankford.edu  
Stanford University, United States of America

### Introduction

This talk explores how new vector-based approaches to computational semantics both afford new methods to digital humanities research, and raise interesting questions for eighteenth-century literary studies in particular. New semantic models known as “word embedding models” have generated excitement recently in the natural language processing and machine learning communities, due to their ability to represent and predict semantic relationships as complex as analogy. “Man” is to “woman” as “king” is to what?, one can ask of the model; “queen.” It will most likely reply. These models formulate analogical and other semantic relationships by computing mathematical vectors for words, such that if  $V(x)$  denotes the vector for the word  $x$ , then the above analogy can be expressed as  $V(woman) - V(man) + V(king) = V(queen)$ . Although these models have a longer history—vector space semantics dates from the ‘70s, having been first developed for the SMART information retrieval system (Salton, 1971) by Gerard Salton and his colleagues (Salton et al., 1975)\* (Turney and Pantel, 2010)—new innovations in their speed and accuracy (see Note [1]) have renewed researchers’ interests—a development begun, in part, by Google, when researchers there unveiled newly efficient algorithms in 2013, packaged in software they released called word2vec. (The word2vec algorithm was originally described by Mikolov et al. 2013. It introduced the neural network to vector space semantics, providing an efficient means by which to compute word vectors. The GloVe algorithm from the Stanford NLP Group eschews the neural network approach, instead performing a novel method of dimensionality reduction on word collocation counts).

“Word vectors,” as these new methods are sometimes informally called, have already enabled published research into questions relevant to humanistic research, such as a recent landmark paper from researchers in the Stanford NLP Group into patterns of

semantic change across centuries of discourse (Hamilton et al.). However, unfortunately, word vectors have so far rarely appeared in research from the digital humanities community itself. Moreover, what work that does exist has so far been primarily circulated through blogs, rather than through published proceedings or articles. Ben Schmidt, for instance, has written an influential introduction to word vectors in his blog post “Vector Space Models for the Digital Humanities” (2015a), which also includes a documented R package for computing them. Also notable is his post, “Rejecting the gender binary” (Schmidt, 2015b), which uses word vectors to dissect the polysemy of words, as well as Michael Gavin’s post, “The Arithmetic of Concepts” (2015), which explores the conceptual implications of adding and subtracting word vectors.

On the whole, the current research landscape of word vectors in the digital humanities resembles the landscape of topic modeling years ago, when the original LDA algorithm (published in 2003 [Blei et al.]), before appearing in landmark published DH studies such as Matt Jockers’ *Macroanalysis* (2013), was employed for humanistic research as early as 2006 by researchers working outside or tangentially to the digital humanities (Newman and Black).

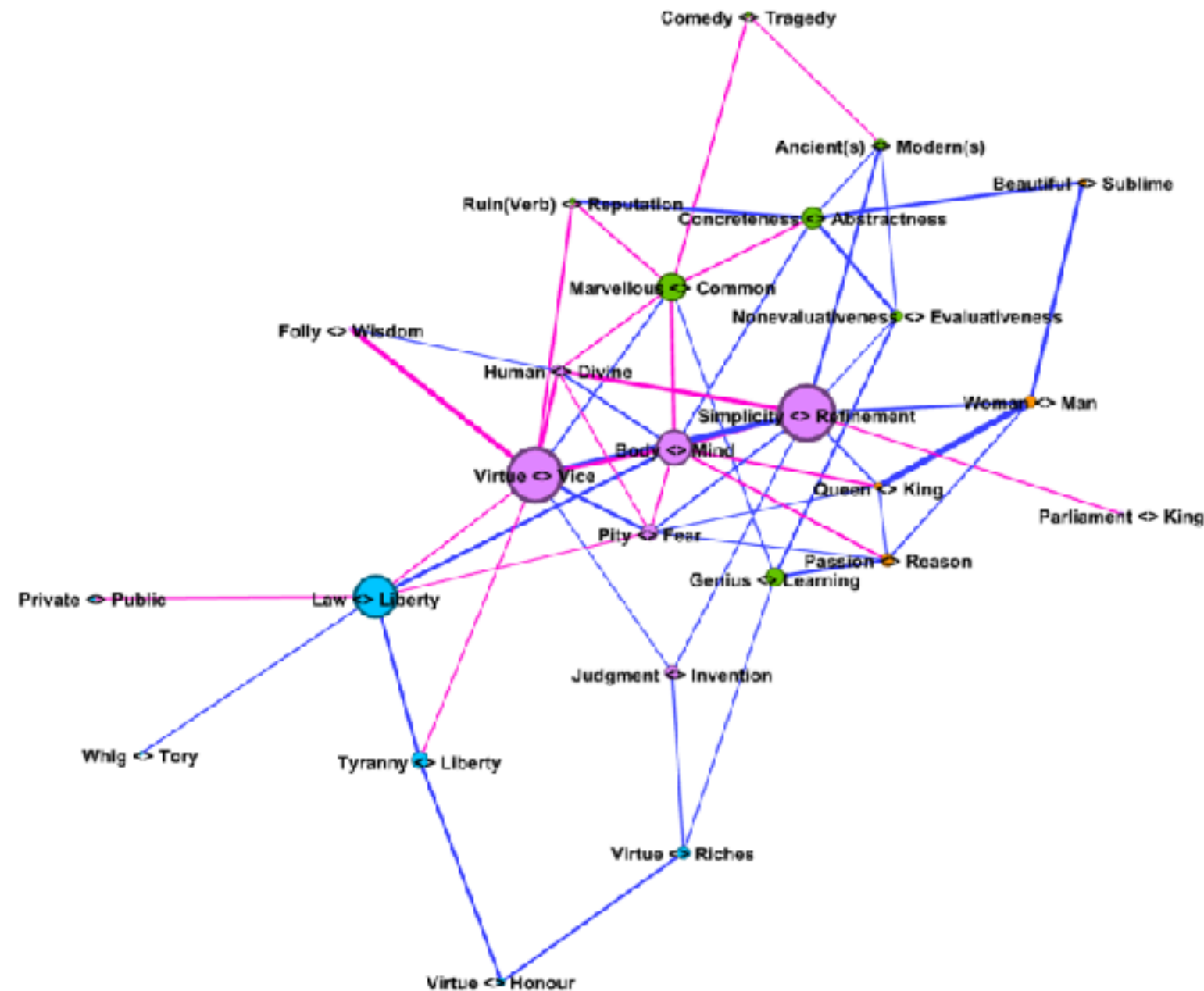
Given this scarcity of digital-humanities research on word vectors, work that seeks equally to explain, interpret, and demonstrate their potential seems particularly useful. With these goals in mind, this paper attempts first to unpack for a digital humanities audience how word vectors work, with reference to the canonical analogy cited above: “man is to woman as king is to queen.” Second, in order to interpret word vectors’ conceptual implications for eighteenth-century literature, I move away from this canonical analogy to one central to a particularly influential argument in the period: “Learning is to Genius as Riches are to Virtue.” Lastly, I turn from this close reading of word vectors to methods of distant-reading analogies that lie implicit in eighteenth-century literature.

### Explaining Word Vectors

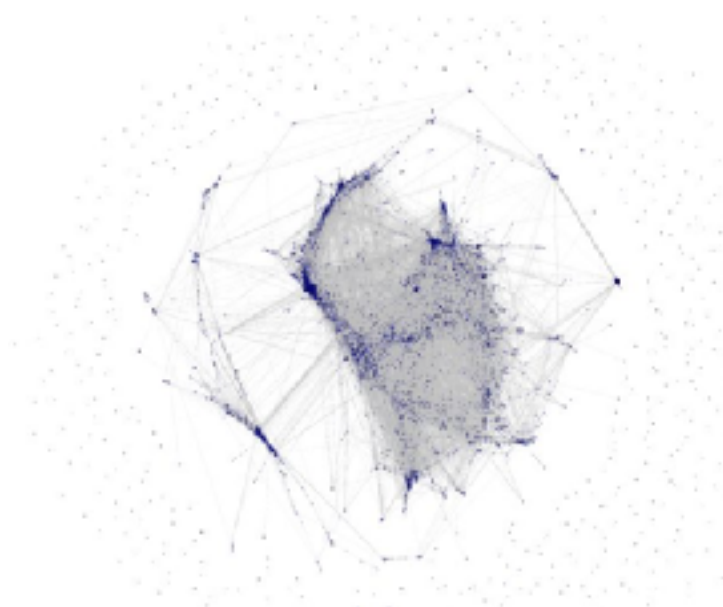
How do word vectors work? In the interests of space, I have omitted this section of my talk from the abstract. Readers curious about the mechanics of word vectors can read more on my blog, which also links to a number of other explanatory resources (Heuser, “Methods”).

### Close-reading Word Vectors

Word vectors provide a persuasive computational means for the semantic representation and analysis of analogies. They combine a mathematical elegance







[meanings](#) | [networks](#) | [spaces](#)

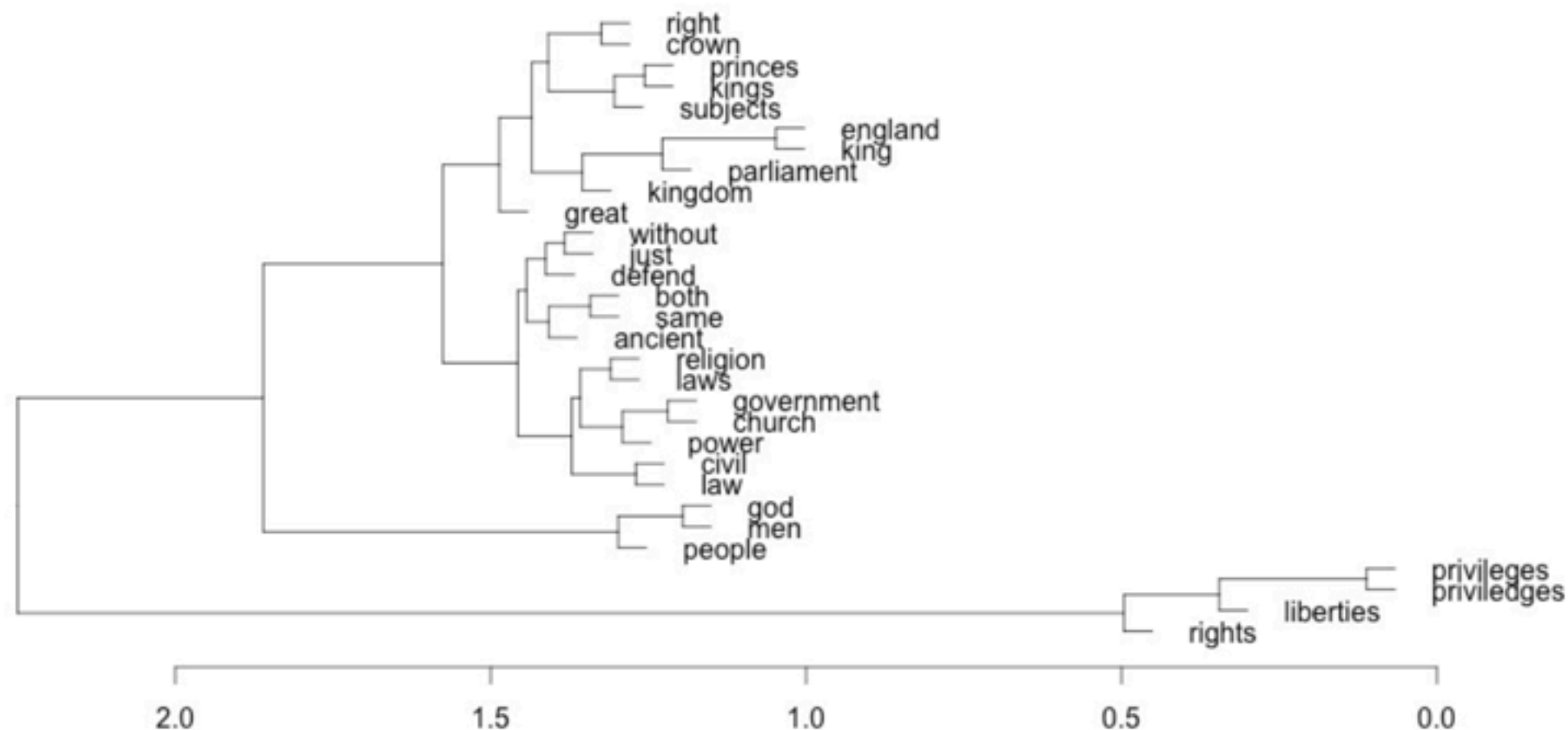
michael gavin, university of south carolina

## Using vector adaptation to read Locke's 'rights'

One common technique for word-sense discrimination uses vector arithmetic (more conventionally called *vector adaptation*) to identify the distinct sense with which a complex word is being used at any given time. Although computers can't tell the difference between 'plane' and 'plane', they can easily tell the difference between ( **pilot** + **plane** + **landing** ) and ( **line** + **plane** + **angle** ). This technique combines a keyword with its surrounding context words, using either component-wise addition or multiplication, thus generating a pseudo-word that reflects their collective place in a semantic network.

To reiterate a point I made too briefly above, I'll argue that this technique can be used to identify two kinds of conceptual implicatures. First, addition lumps a keyword together with all its context words, capturing every possible connection in the network. My sense is that these baggy pseudo-words, when compared to real words across the corpus, will tend to point to the most important, broadly conceived underlying assumptions that inform any statement. They are particularly good at identifying implicit binaries that operate as organizing contrasts. Component-wise addition gives us something close to the axiomatic, noetic, or load-bearing concepts (to use de Bolla's terminology) that support a given statement.

### rights ( sig )





word2vec

Interestingly, **word-embedding models** can mathematically represent and predict semantic relationships between words as complex as **analogy**.



---

## Efficient Estimation of Word Representations in Vector Space

---

**Tomas Mikolov**

Google Inc., Mountain View, CA  
tmikolov@google.com

**Kai Chen**

Google Inc., Mountain View, CA  
kaichen@google.com

**Greg Corrado**

Google Inc., Mountain View, CA  
gcorrado@google.com

**Jeffrey Dean**

Google Inc., Mountain View, CA  
jeff@google.com

### Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

## 1 Introduction

Many current NLP systems and techniques treat words as atomic units - there is no notion of similarity between words, as these are represented as indices in a vocabulary. This choice has several good reasons - simplicity, robustness and the observation that simple models trained on huge amounts of data outperform complex systems trained on less data. An example is the popular N-gram model used for statistical language modeling - today, it is possible to train N-grams on virtually all available data (trillions of words [3]).

However, the simple techniques are at their limits in many tasks. For example, the amount of relevant in-domain data for automatic speech recognition is limited - the performance is usually dominated by the size of high quality transcribed speech data (often just millions of words). In machine translation, the existing corpora for many languages contain only a few billions of words or less. Thus, there are situations where simple scaling up of the basic techniques will not result in any significant progress, and we have to focus on more advanced techniques.

With progress of machine learning techniques in recent years, it has become possible to train more complex models on much larger data set, and they typically outperform the simple models. Probably the most successful concept is to use distributed representations of words [10]. For example, neural network based language models significantly outperform N-gram models [1, 27, 17].

### 1.1 Goals of the Paper

The main goal of this paper is to introduce techniques that can be used for learning high-quality word vectors from huge data sets with billions of words, and with millions of words in the vocabulary. As far as we know, none of the previously proposed architectures has been successfully trained on more

"What is the word that is similar to *small* in the same sense as *biggest* is similar to *big*?"



---

## Efficient Estimation of Word Representations in Vector Space

---

Tomas Mikolov  
Google Inc., Mountain View, CA  
tmikolov@google.com

Kai Chen  
Google Inc., Mountain View, CA  
kaichen@google.com

Greg Corrado  
Google Inc., Mountain View, CA  
gcorrado@google.com

Jeffrey Dean  
Google Inc., Mountain View, CA  
jeff@google.com

### Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

### 1 Introduction

Many current NLP systems and techniques treat words as atomic units - there is no notion of similarity between words, as these are represented as indices in a vocabulary. This choice has several good reasons - simplicity, robustness and the observation that simple models trained on huge amounts of data outperform complex systems trained on less data. An example is the popular N-gram model used for statistical language modeling - today, it is possible to train N-grams on virtually all available data (trillions of words [3]).

However, the simple techniques are at their limits in many tasks. For example, the amount of relevant in-domain data for automatic speech recognition is limited - the performance is usually dominated by the size of high quality transcribed speech data (often just millions of words). In machine translation, the existing corpora for many languages contain only a few billions of words or less. Thus, there are situations where simple scaling up of the basic techniques will not result in any significant progress, and we have to focus on more advanced techniques.

With progress of machine learning techniques in recent years, it has become possible to train more complex models on much larger data set, and they typically outperform the simple models. Probably the most successful concept is to use distributed representations of words [10]. For example, neural network based language models significantly outperform N-gram models [1, 27, 17].

#### 1.1 Goals of the Paper

The main goal of this paper is to introduce techniques that can be used for learning high-quality word vectors from huge data sets with billions of words, and with millions of words in the vocabulary. As far as we know, none of the previously proposed architectures has been successfully trained on more

"Somewhat surprisingly, these questions can be answered by performing **simple algebraic operations with the vector representation of words**. To find a word that is similar to *small* in the same sense as *biggest* is similar to *big*, we can simply compute vector  $X = \text{vector}(\text{"biggest"}) - \text{vector}(\text{"big"}) + \text{vector}(\text{"small"})$ . Then, we search in the vector space for the word closest to  $X$  measured by cosine distance. When the word vectors are well trained, it is possible to find the correct answer (word *smallest*) using this method."



Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks



Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Which more **abstract concepts** can be detected through word-embedding models? Is it possible to discern **the conceptual limits of what can be said and stated (Foucault)** in a larger corpus, for example?



The background is a historical manuscript page, likely from a technical or scientific work. At the top, there is a title in Swedish: "Den första Machinen om Wallinhs Kraft på ett fall." (The first machine about Wallin's power on a fall). Below the title are several small diagrams of mechanical components, including wheels and shafts, labeled with letters A, B, C, D, E, F, G, H, I, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z. On the right side, there is a large, detailed drawing of a machine, possibly a pump or a mill, with a frame and various mechanical parts. The text "N: 1." is written in the top right corner. The overall style is that of an 18th-century technical drawing.

1. Introduction

**2. From Archival to Data Driven Humanistic Research**

3. About the Research Project "Digital Models"

4. Textual Models of the Past

5. Visual Models of the Past

6. Conclusion

All slides in the form of a PDF can be found at <http://pellesnickars.se/>



**Images from Nyköping (Svenska Bio, 1909)**



# 1897

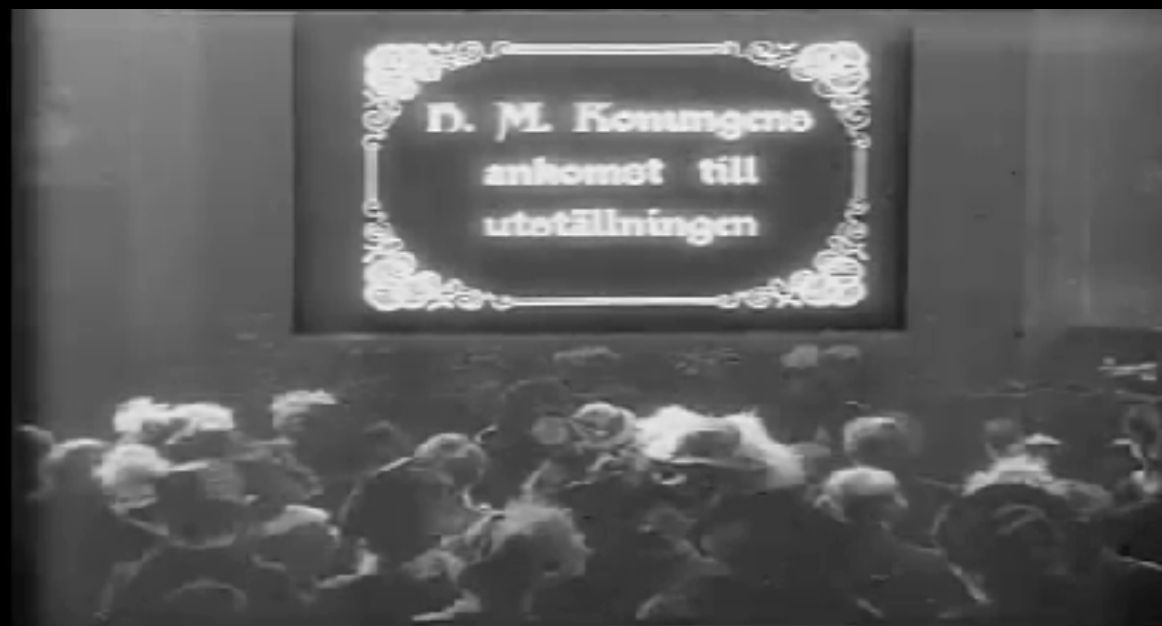
Mediehistorier kring  
Stockholmsutställningen

RED. ANDERS EKSTRÖM,  
SOLVEIG JÜLICH & PELLE SNICKARS



GAMLA S

D. M. Konungens  
ankomst till  
utställningen





# 1897

## Mediehistorier kring Stockholmsutställningen

RED. ANDERS EKSTRÖM,  
SOLVEIG JÜLICH & PELLE SNICKARS



GAMLA S

Pelle Snickars

### Mediearkeologi: Om utställningen som mediearkiv

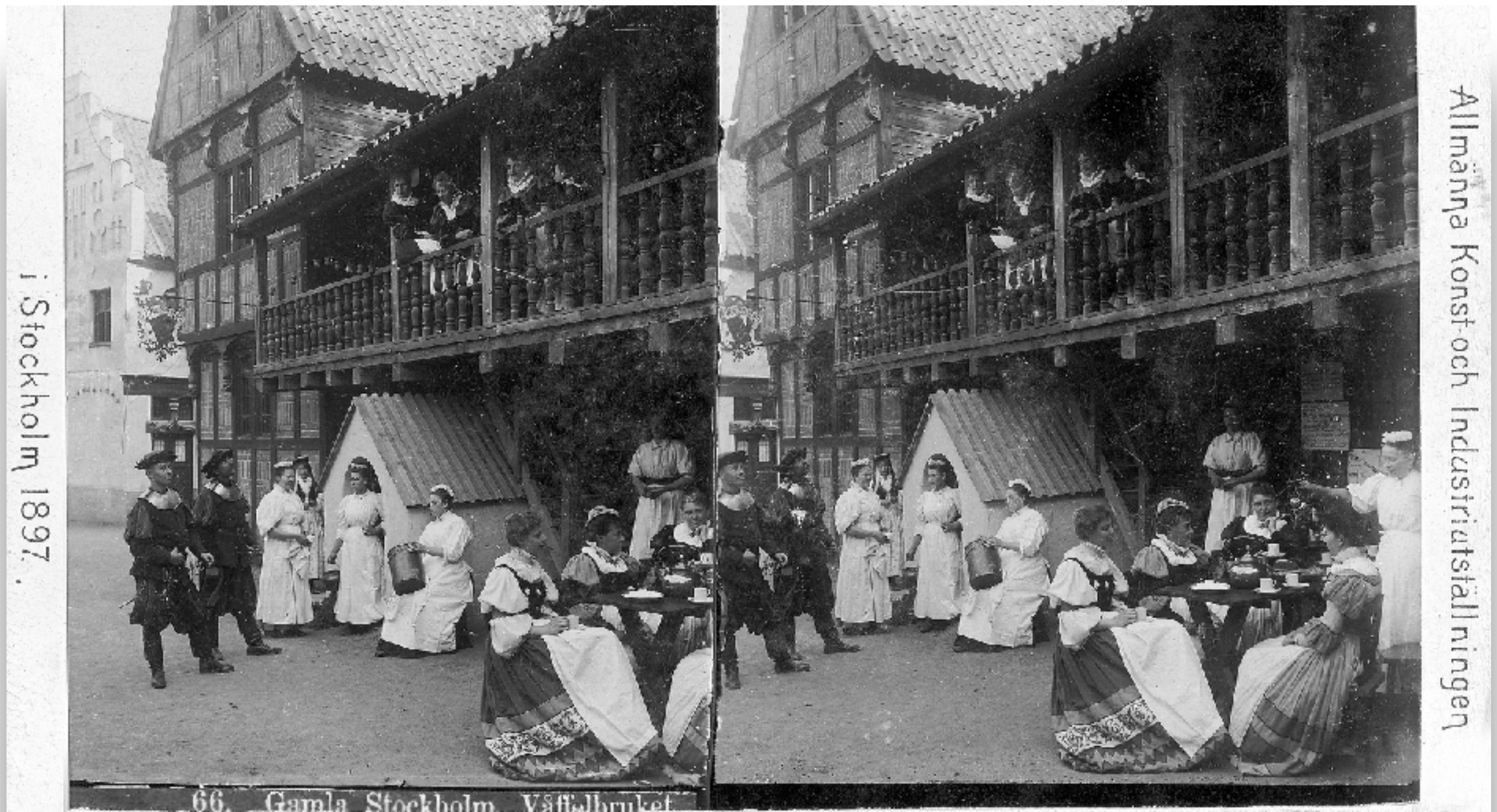
I slutet av Andreas Hasselgrens krönika över Stockholmsutställningen 1897 finns ett vackert fotografi med titeln "Vy af utställningsområdet. (Den 15 mars 1898)".<sup>1</sup> På bilden är utställningsområdet höjt i snö och inte en människa är synlig. Men något annat hade varit förvånande. När översiktsskildern togs uppe från Skansens sluttningar med det halvfärdiga Nordiska museet och Strandvägen i bildfonden, hade den Allmänna konst- och industriutställningen varit stängd i mer än ett halvår. Platsen där den ägt rum på Lejonslätten på Djurgården var inte bara folktom. Alla de byggnader och paviljonger som tidigare varit uppställda på området var också borta.



Ett halvår efter stängningen fanns bara spår och lämningar av utställningen kvar. Fotografi av Axel Lindahls fotografiaffär. Ur Andreas Hasselgren, *Utställningen i Stockholm 1897: Beskrifning i ord och bild öfver Allmänna konst- & industriutställningen* (Stockholm, 1897).



# Archival driven media historical cultural research ...







[Blogg](#) / [Böcker](#) / [Redaktionsråd](#)  
/ [Om bokserien](#) / [Manus](#) /  
[Kontakt](#)

#### Om bokserien

Mediehistoriskt arkiv publicerar forskare från hela Sverige och rymmer antologier, monografier – inklusive avhandlingar – och källsamlingar. Serien ges ut av ämnet mediehistoria vid Institutionen för kommunikation och medier vid Lunds universitet. Alla böcker är CC-licensierade – erkännande, icke-kommersiell, inga bearbetningar 3.0. Vi ser gärna att böckerna sprids och används så mycket som möjligt.



#### MEDIEHISTORISKT ARKIV 28

##### Återkopplingar

Marie Cronqvist, Patrik Lundell, Pelle Snickars (red.), 2014

Det bedrivs alltför lite mediehistorisk forskning i Sverige. Mediehistoria kan – och bör – skrivas på många olika sätt. En ambition inom den kulturhistoriska medieforskning som presenteras i den kommande boken. Återkopplingar, är att genom ett breddat mediebegrepp och historisk sensibilitet uppdatera mediestudiet. Förnyelsen sker inte sällan i skärningspunkten mellan den ofta teknikdeterministiska mediearkeologin och...

[Läs mer](#)



#### MEDIEHISTORISKT ARKIV 26

##### Information som problem: Medieanalytiska texter från medeltid till framtid

Olufried Czulka, Jonas Nordin & Pelle Snickars (red.), 2014

Vad är information? Är information det samma som kunskap? Har värderingen av information förändrats över tid? Är all information viktig? Är all information nyttig? Vem kontrollerar informationen? Skall information kontrolleras? Kan information vara fri?



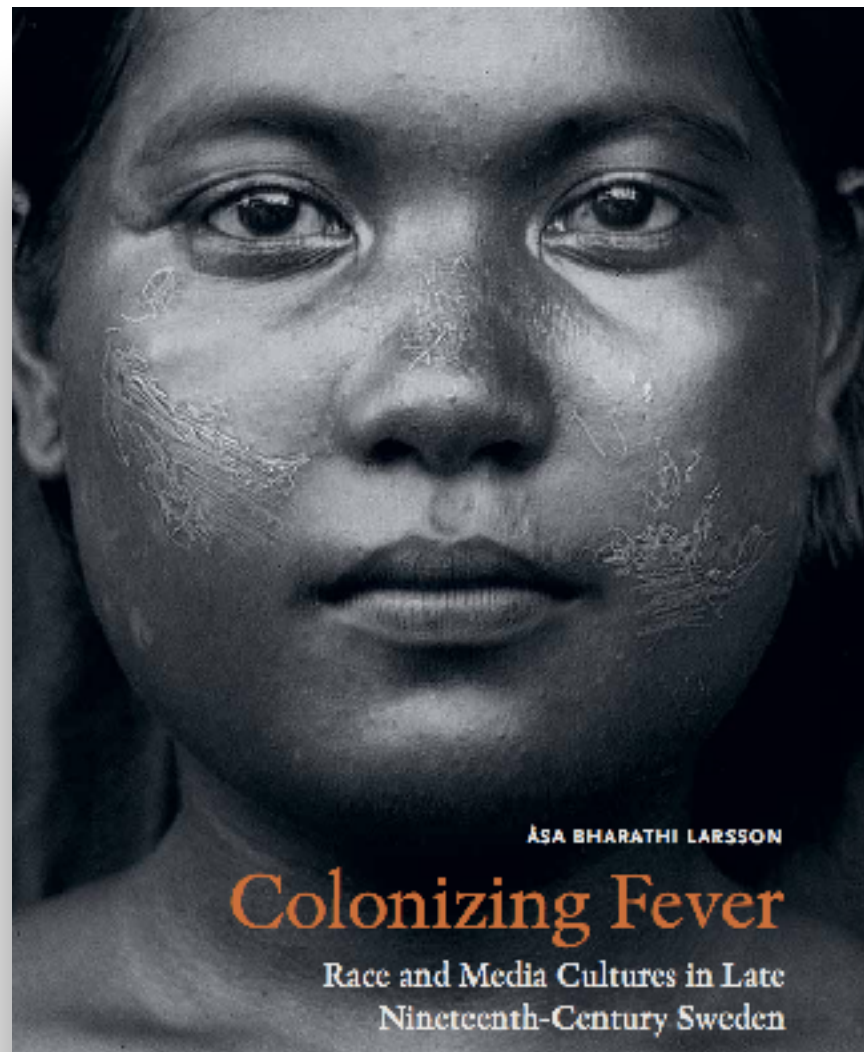
#### MEDIEHISTORISKT ARKIV 25

##### Mediehistoriska vändningar

Marie Cronqvist, Johan Jarlbrink, Patrik Lundell (red.), 2014

Sedan sekelskiftet 2000 har det talats om en medial vändning inom framför allt humanvetenskaperna. För till exempel litteraturvetenskapens del har det handlat om en fokusörflyttning från estetiken till historiska medieanalyser av sådant som läs- och skrivpraktiker, och motsvarande intresseförskjutningar kan avläsas inom andra discipliner. I en svensk kontext har det rentav hävdats att medie- och... [Läs mer](#)



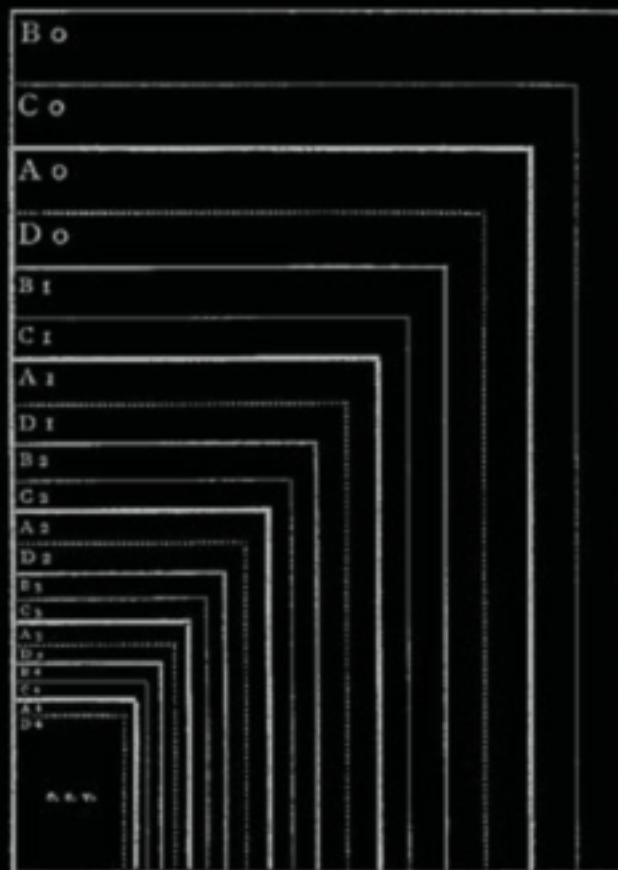




# Pappersarbete

FORMANDET AV OCH FÖRESTÄLLNINGAR OM  
KONTORSPAPPER SOM MEDIUM

Charlie Järpevall



Mats Lindström | Drömmar om det minsta  
MICROFILM, ÖVERFLÖD OCH BRIST  
1900-1970

















digitisation



From **documents to data**—a shift that concerns the **epistemic foundation and status of the library and “the archive”** (understood in an abstract Foucauldian sense as a system of discursivity that establishes the limits of what can be said and stated).







## DIGITAL HUMANITIES (DIGIHUM)

ACADEMY PROGRAMME 2016–2019

Programme memorandum





## DIGITAL HUMANITIES (DIGIHUM)

ACADEMY PROGRAMME 2016–2019

Programme memorandum



Vetenskapsrådet



### Idéseminarium om digitalisering och tillgängliggörande av kulturarvssamlingar

Vetenskapsrådet har fått regeringens uppdrag att stödja data driven forskning. Halva denna betydande satsning rör digitalisering och tillgängliggörande av kulturarvssamlingar, främst inom humaniora och samhällsvetenskap, men även inom andra vetenskapliga områden med högt värde för forskning<sup>1</sup>. Riksbankens jubileumsfond och Kungliga Vitterhetsakademien har tidigare gjort gemensamma satsningar inom detta område. Tillsammans planerar nu Vetenskapsrådet, Riksbankens jubileumsfond och Kungliga Vitterhetsakademien hur fortsatta satsningar bör göras för att bäst svara mot forskningens behov. Vi bjuder därför in forskare samt företrädare för kulturarvsinstitutioner till ett idéseminarium. Syftet är att få underlag för utformning av eventuella utlysningar eller andra gemensamma satsningar under kommande år för att skapa en ökad tillgång till stora digitala datavolymer som kan öppna upp för nya frågeställningar och metodutvecklingar.

Seminariet hålls den 17 oktober 2017 kl. 13.00 – 17.00 i Kreugersalen, Tändstickspalatset, Västra Trädgårdsgatan 15 i Stockholm. Från kl. 12.30 serveras en enkel smörgåslunch.

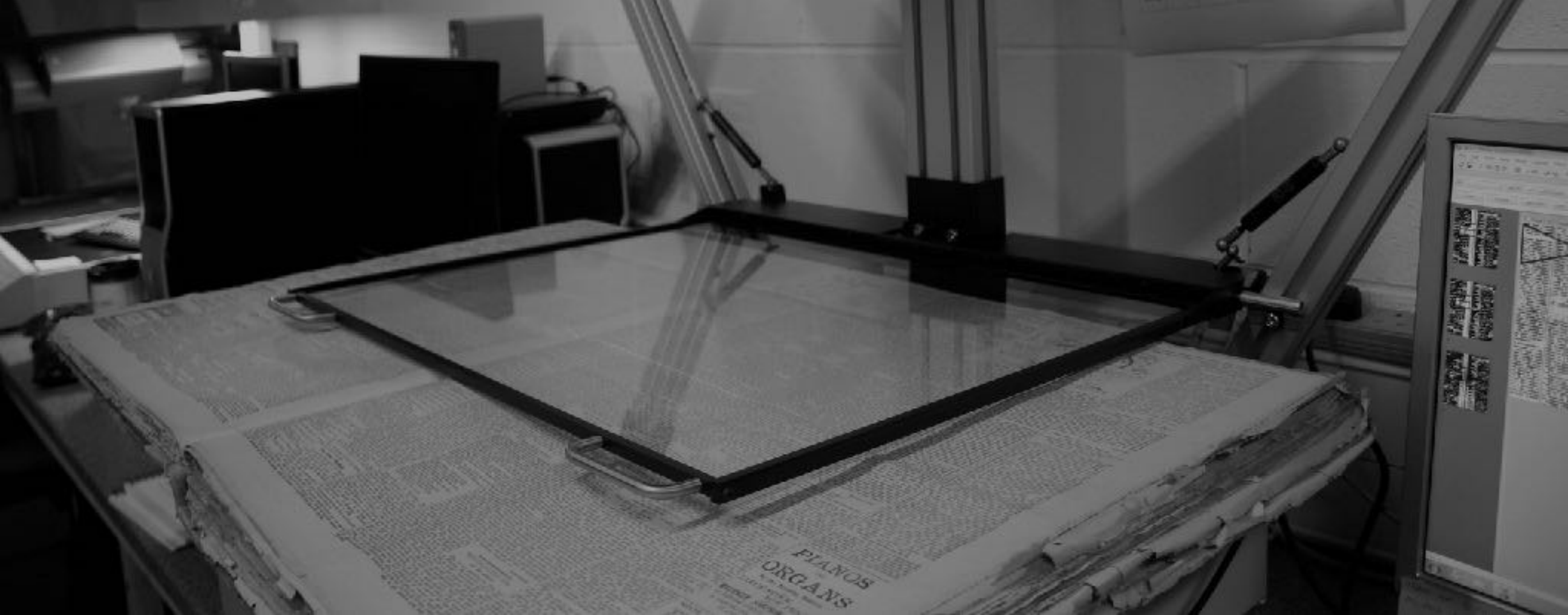
Vi ser fram emot att få ta del av synpunkter, erfarenhet och förslag från såväl talare som seminariedeltagare.

### Hjärtligt välkomna!

Kerstin Sahlin, Huvudsekreterare för Humaniora och samhällsvetenskap vid Vetenskapsrådet  
Göran Blomqvist, VD för Riksbankens jubileumsfond  
Birgitta Svensson, ständigt sekreterare för Kungliga Vitterhetsakademien

<sup>1</sup> Utdrag ur Vetenskapsrådets regleringsbrev för 2017: Vetenskapsrådet ska förbereda för att finansiera dels digital infrastruktur som stöder data driven forskning, dels digitalisering och tillgängliggörande av kulturarvssamlingar, främst inom humaniora och samhällsvetenskap, men även inom andra vetenskapliga områden med högt värde för forskning. Avseende tillgängliggörande av kulturarvssamlingar ska Vetenskapsrådet samverka med Samordningssekreteraria för digitalisering, digitalt bevarande och digitalt tillgängliggörande av kulturarvet (Digisam) vid Riksantikvarieämbetet.

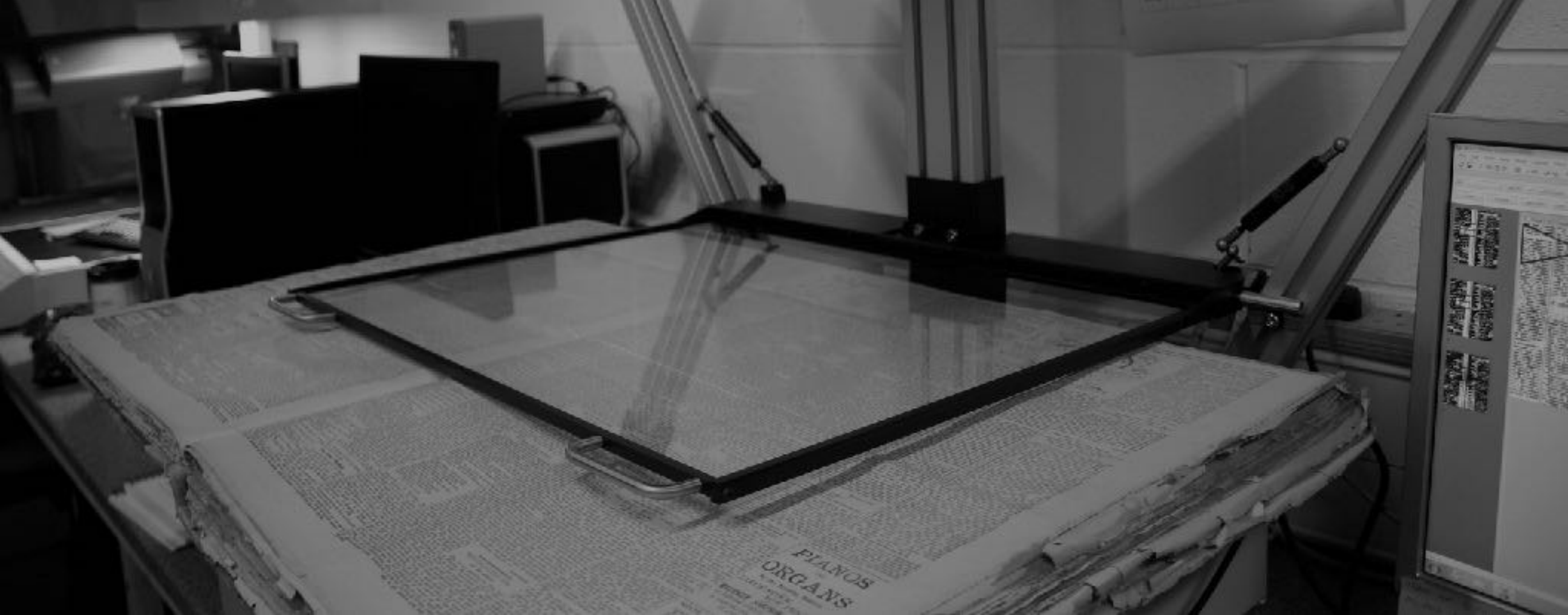




From my perspective **data driven humanities research**—with a focus on the historical sciences—departs from the fact the epistemic foundation of “the archive” has been **altered and modified due to digitisation**.

Nothing more, nothing less.






As a consequence, some humanistic research practices—but far from all—needs to adapt and try **to take advantage of scholarly possibilities** that arise when **'documents as data'** are **sharable and networked, linkable and traceable, reusable and processable.**



# LOS ANGELES REVIEW OF BOOKS

REVIEWS ESSAYS INTERVIEWS SECTIONS BLARB PRINT AV ABOUT DONATE



## Neoliberal Tools (and Archives): A Political History of Digital Humanities

By Daniel Allington, Sarah Brouillette, David Golumbia

8670

0

22



MAY 1, 2016



# The Digital-Humanities Bust

After a decade of investment and hype, what has the field accomplished? Not much.

By Timothy Brennan | OCTOBER 15, 2017

✓ PREMIUM



Andrea Levy for The Chronicle Review

First came the debacle of the high-priced "Ada" algorithm, the control center of Hillary Clinton's ill-fated operation. Next ESPN wonk Nate Silver, after flubbing the 2016 election forecast, defended his numbers by claiming that he was not more wrong than every other statistical predictor since 1968. Finally, consider the kerfuffle over Cambridge Analytica, the British company whose "psychographics" method of data modeling and "emotion analysis" claimed to be the Trump camp's secret weapon — until skeptics recalled that Ted Cruz and Ben Carson had employed their services as well.

The dream that algorithmic computation might reveal the secrets of complex social and cultural processes has suffered a very public and embarrassing results crisis. These setbacks have also led to some soul-searching in the university, prompting a closer look at the digital humanities. Roughly a decade's worth of resources have now been thrown in their direction, including



## Data Driven Humanistic Research

What **type of data** is going to drive the research process?



## Data Driven Humanistic Research

Well, it naturally depends and hinges on which research question that are being posed—in **analogy to the ways in which historical sources** are used and conceptualised within the various sciences of history.





## British Library Data Strategy 2017





“Data are evidence that underpins the answer to the research question, and can be used to validate findings regardless of its form (e.g. print, digital, or physical).”



## From Archival to Data Driven Humanistic Research

My own research was for a number of years driven by what I **found in "the archive"**.



## From Archival to Data Driven Humanistic Research

My own research was for a number of years driven by what I **found in "the archive"**.

Today—given the rapid pace of the **digitisation of cultural heritage**—it is increasingly driven by the access and analyses of data, including the **various tools and methods** on offer.



## From Archival to Data Driven Humanistic Research

My own research was for a number of years driven by what I **found in "the archive"**.

Today—given the rapid pace of the **digitisation of cultural heritage**—it is increasingly driven by the access and analyses of data, including the **various tools and methods** on offer.



## From Archival to Data Driven Humanistic Research

Importantly, data driven humanities research **needs to pay attention** to the **altered materiality of the archive**, the **production processes of data**, to various forms of **data "agency"** or the **epistemic machinery** that produces data etcetera etcetera.



Workshop datadriven humanistisk forskning 15/9  
2017, KTH

patriksv August 21, 2017 - 4:51 pm

## **datadriven forskning**

snabb takt kraftigt ökande  
e sätt förändrat förutsättnin

**Workshop om datadriven humanistisk forskning**  
**15 september 2017, KTH (#DHUM17)**  
**A108, Arkitektskolan**

Denna sida innehåller program, inbjudna deltagare och referenser/material för den workshop om datadriven forskning i humaniora som äger rum på KTH, Stockholm, den 15 september 2017. Organisatörer är Pelle Snickars, Umeå universitet och Patrik Svensson, Umeå universitet/UCLA.



“Digitally driven humanistic research has in recent years shown a potential to connect critical and technological perspective—where the task is not only about managing data, building tools or studying digitally distant phenomena but also to engage and intervene with this data.”

<http://patriksv.net/2017/09/om-datadriven-forskning/>





1. Introduction

2. From Archival to Data Driven Humanistic Research

**3. About the Research Project "Digital Models"**

4. Textual Models of the Past

5. Visual Models of the Past

6. Conclusion

All slides in the form of a PDF can be found at <http://pellesnickars.se/>

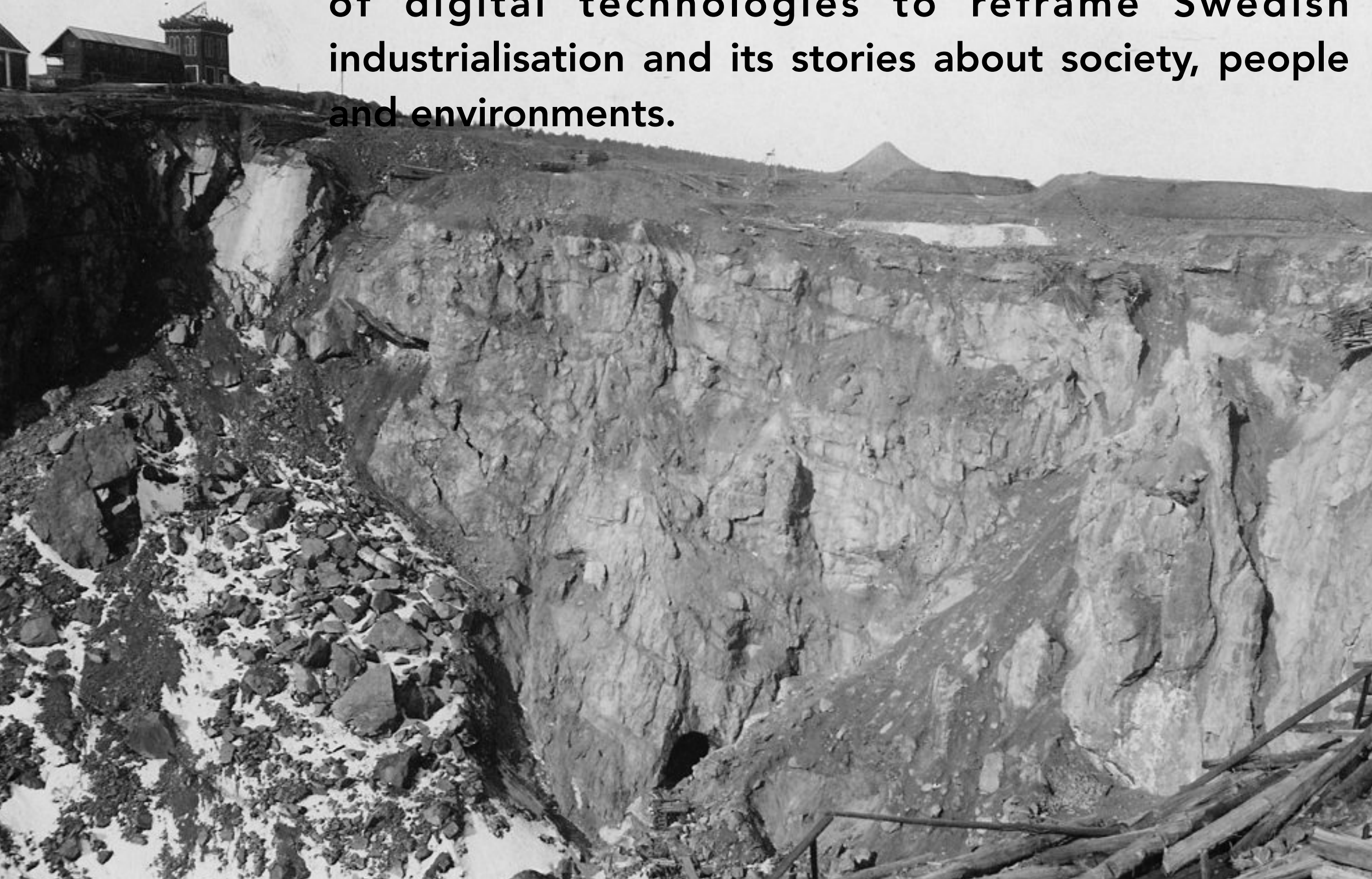




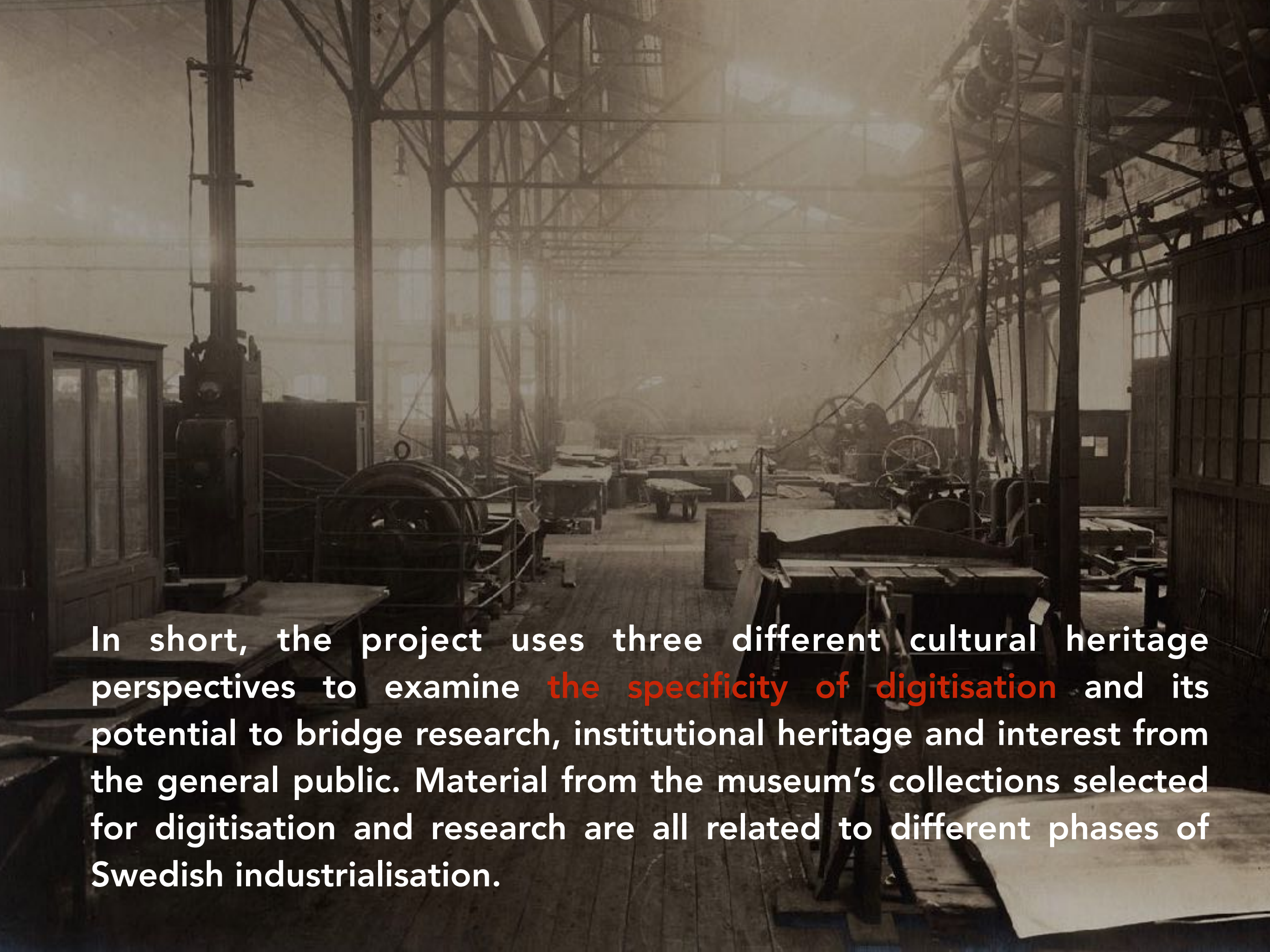
The research project "Digital Models. Techno-historical collections, digital humanities & narratives of industrialisation" is funded by the Royal Swedish Academy of Letters, History and Antiquities between 2016 and 2019. It is a collaboration between the Swedish National Museum of Science and Technology—located in Stockholm and with a national responsibility Sweden's technical and industrial heritage—and the digital humanities hub, Humlab at Umeå University.



Based on selected parts of the Technical museum's collections the project aims to explore the potential of digital technologies to reframe Swedish industrialisation and its stories about society, people and environments.







In short, the project uses three different cultural heritage perspectives to examine **the specificity of digitisation** and its potential to bridge research, institutional heritage and interest from the general public. Material from the museum's collections selected for digitisation and research are all related to different phases of Swedish industrialisation.

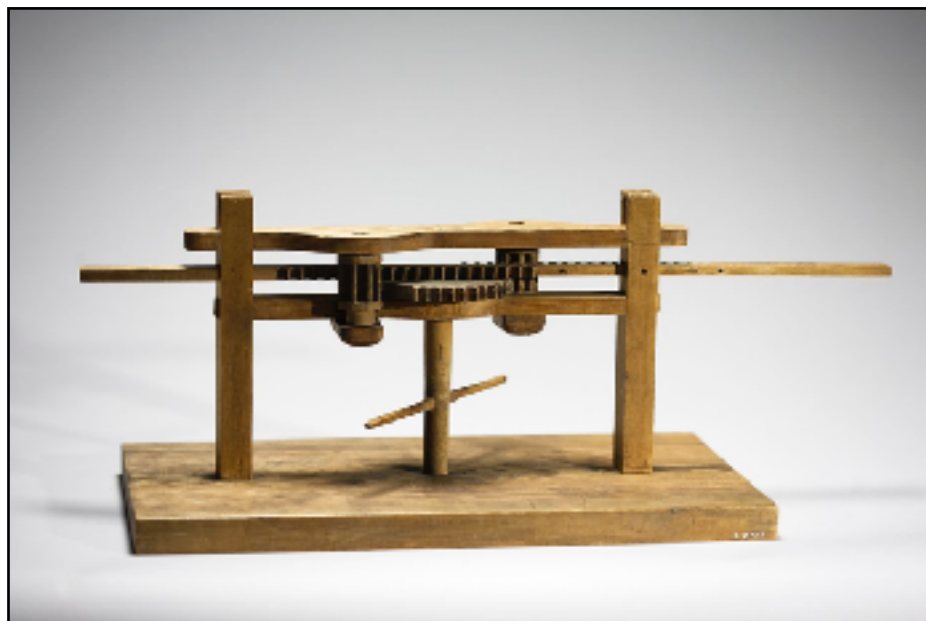




(A). Parts of the business leader and industry historian, Carl Sahlin's (1861-1943) extensive collection.



(B). All editions of the museum yearbook, *Daedalus* ( 1931-2014).



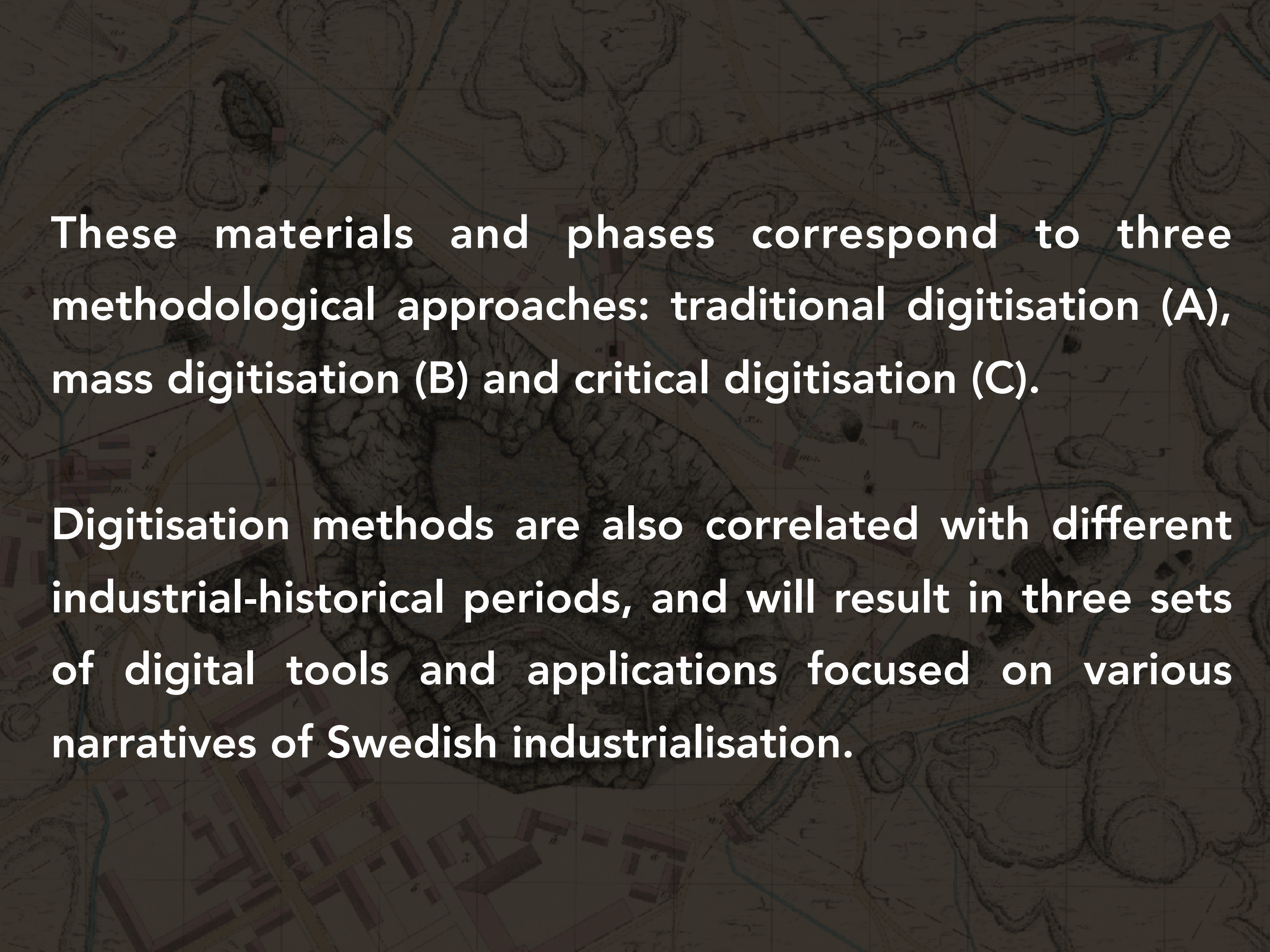
(C). 31 wood models from Swedish pre-industrial inventor Christopher Polhem's "mechanical alphabet" from the early 1700s.



Karta öfver Fahlu eller Stora Kopparbergs Grufwor (1845)







These materials and phases correspond to three methodological approaches: traditional digitisation (A), mass digitisation (B) and critical digitisation (C).

Digitisation methods are also correlated with different industrial-historical periods, and will result in three sets of digital tools and applications focused on various narratives of Swedish industrialisation.





Med utgångspunkt i Tekniska museets samlingar utforskar vi den digitala teknikens möjligheter att omgestalta industrialiseringens berättelser om samhälle, människor och miljöer.



### Modell 1: Sahlins arkiv

Vår bild av industrialismen är fast cementerad i berättelser om framsteg, materiell utveckling och manliga bedrifter. Hur kan digital teknik hjälpa till att finna nya ingångar till befintliga samlingar, samt att nyansera och problematisera bilden av industrialismen?



### Modell 2: Dædalus

Dædalus är en årsbok som alltsedan 1931 har publicerats av Tekniska museet. Projektmodell 2 handlar om att massdigitalisera denna tidskrift och studera dess totala textmängder i jakt efter lingvistiska och teknik-, miljö-, medie- och genushistoriskt signifikanta mönster.



### Modell 3: Polhems alfabet

Kan man utvinna ny historisk kunskap ur Christopher Polhems mekaniska alfabet som digital modell – och samtidigt använda artefakterna för pedagogiska ändamål som svarar mot samtidens behov?



# Digitala modeller

Medverkande & kontakt



**Anders Hultz**

Docent i teknik- och vetenskapshistoria och forskningschef vid Centrum för Näringslivshistoria



**Anna Foka**

Biträdande lektor i humaniora och informationsteknik vid HUMlab, Umeå universitet



**Anna-Karin Nilsson Stål**

Intendent på Tekniska museets avdelning Kunskap & Samlingar



**Åsa Marnell**

Avdelningschef för Kunskap & Samlingar på Tekniska museet



**Britta Isaksson-Bergholm**

Pedagog och intendent, Tekniska museet



**Finn Arne Jørgensen**

Docent och universitetslektor i teknik- och miljöhistoria vid Umeå universitet



**Jenny Attemark-Gillgren**



**Lotta Oudhuis**



**Pelle Snickars**



# Karta öfver Fahlu eller Stora Kopparbergs Grufwor (1845)

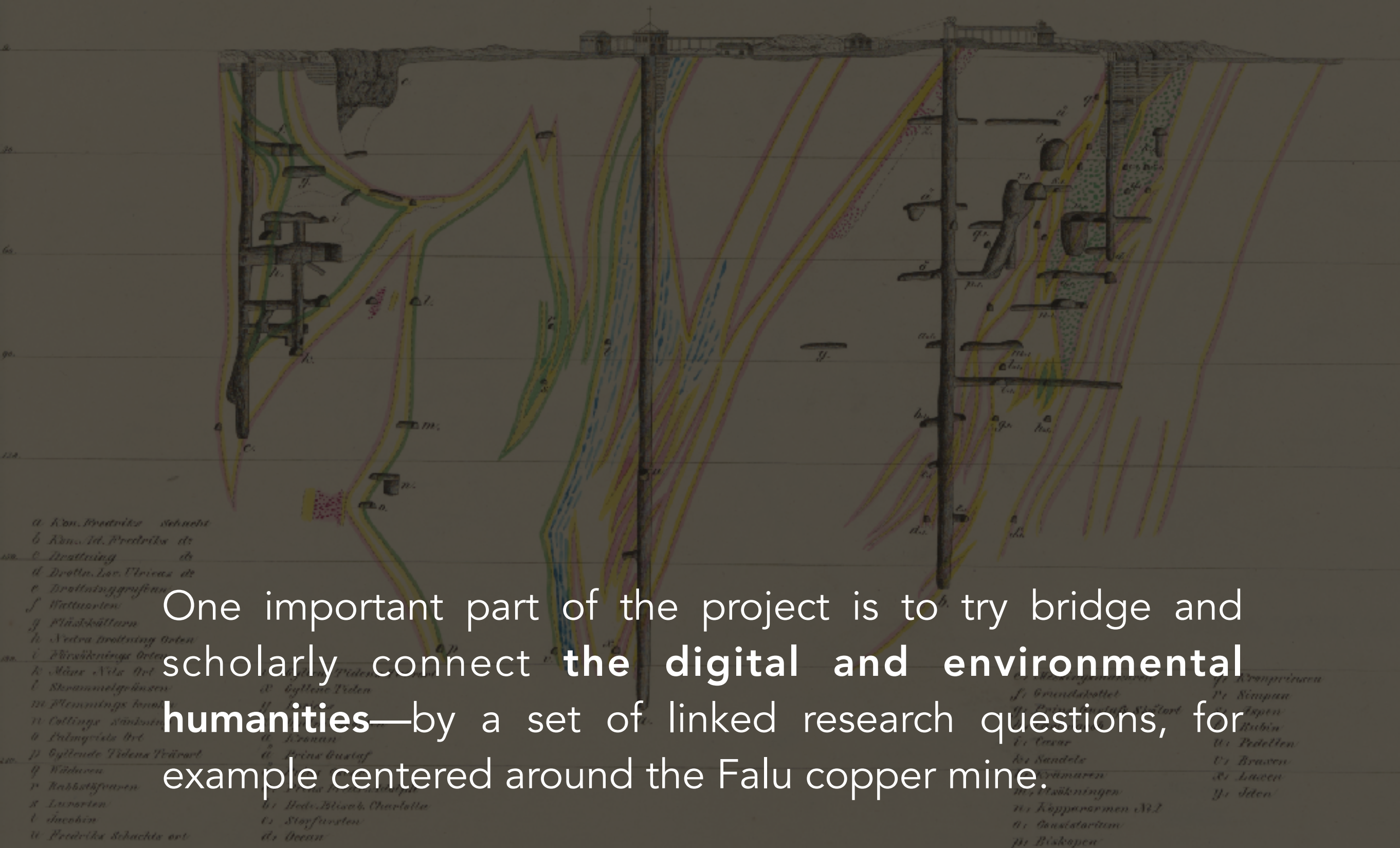
Skärning genom Drätlings Grufwan, Konung Fredriks, Adolph Fredriks och Drottning Lovisa Ulricas Schachter i O.S.O. V.A.V.





# Karta öfver Fahlu eller Stora Kopparbergs Grufwor (1845)

*Skärning genom Drätlings Grufvan, Konung Fredriks, Adolph Fredriks och Drottning Lovisa Ulricas Schachter i O.S.O. V.A.V.*





## Modell 1 (A). Carl Sahlin

# Digitizing Carl Sahlin's archive: A digital environmental humanities perspective

Finn Arne Jørgensen

Professor of environmental history

University of Stavanger, Norway











The DH scholar **Anna Foka** is, for example, interested in the very nature and structure of **the Sahlin archive (A.)**—and if it is possible for new digitisation practices or tools to bring in a feminist perspective on **an archive assembled pre-feminism.**





Is it, for instance, possible with different digitisation practices to look into the Sahlin archive and **retell stories of women in the Swedish industry that are concealed and untold?**



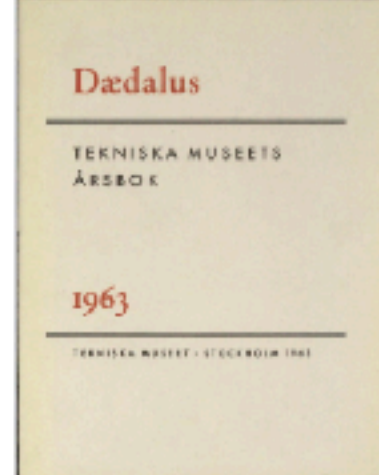
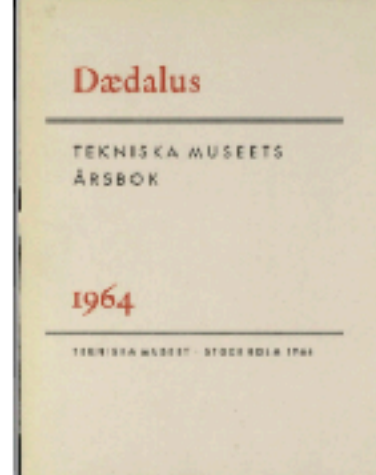
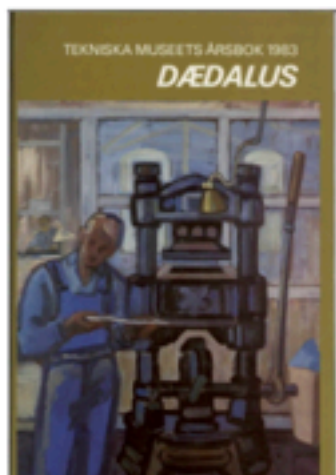
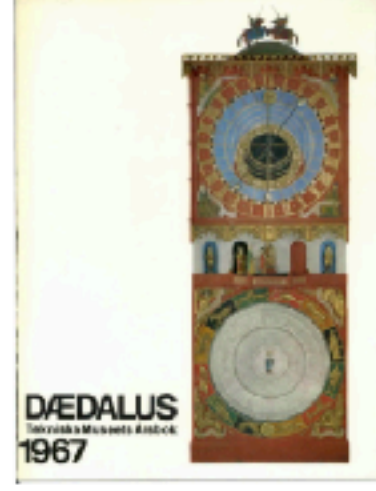
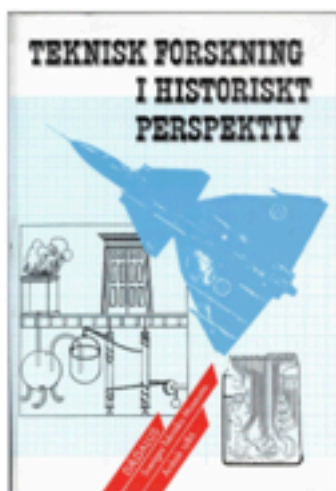
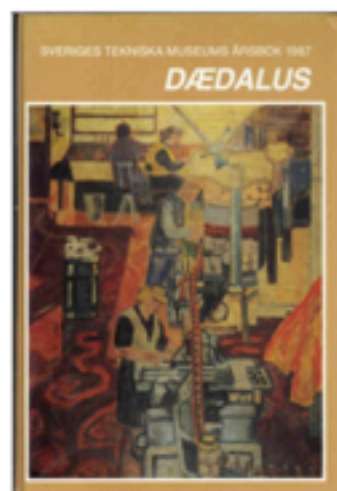
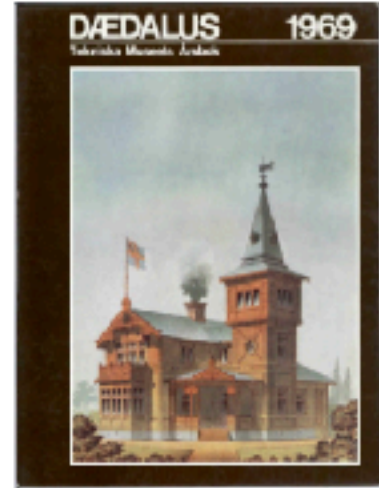
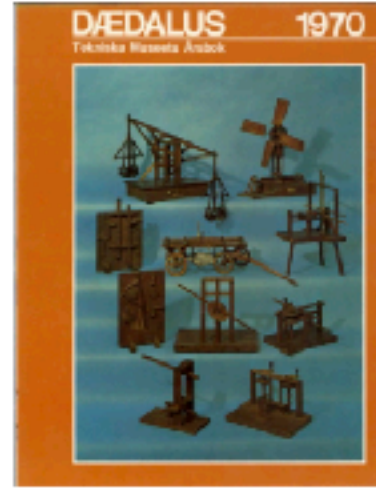
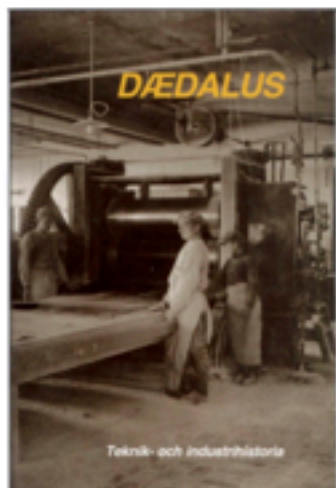
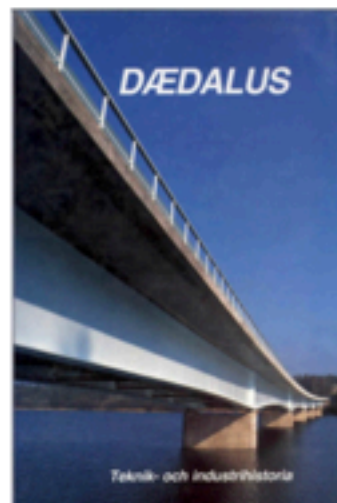
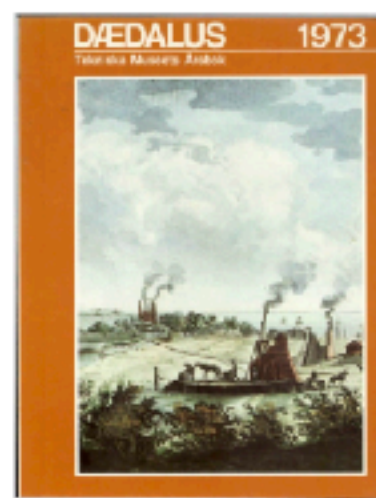
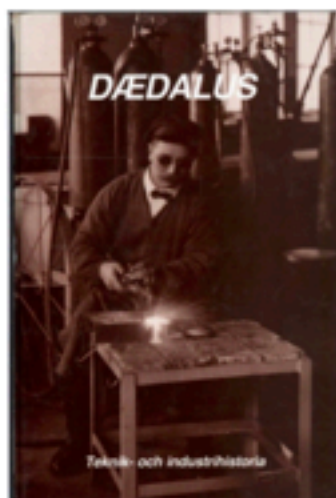


Carl Sahlin's extensive collection at the Technical museum was crucial for the inauguration of the museum.

In a similar vein the wooden models by Polhem were also important objects for the museum—hence the project uses a **meta-museological perspective** of the ways in which digitisation today alters museological practices.









TAL  
OM  
NYTTAN  
AF ET  
*Laboratorium Mechanicum,*  
HÅLLIT FÖR  
KONGL. VETENSKAPS ACADEMIEN  
AF  
CARL KNUTBERG,  
CAPITAINE MECHANICUS,  
DÅ HAN DER BLEF SÅSOM LEDA-  
MOT INTAGEN  
DEN 16 NOVEMBER 1754.



På Kongl. Vetenskaps Academiens befallning.

STOCKHOLM,  
Tryckt hos LARS SALVIUS, 1754.

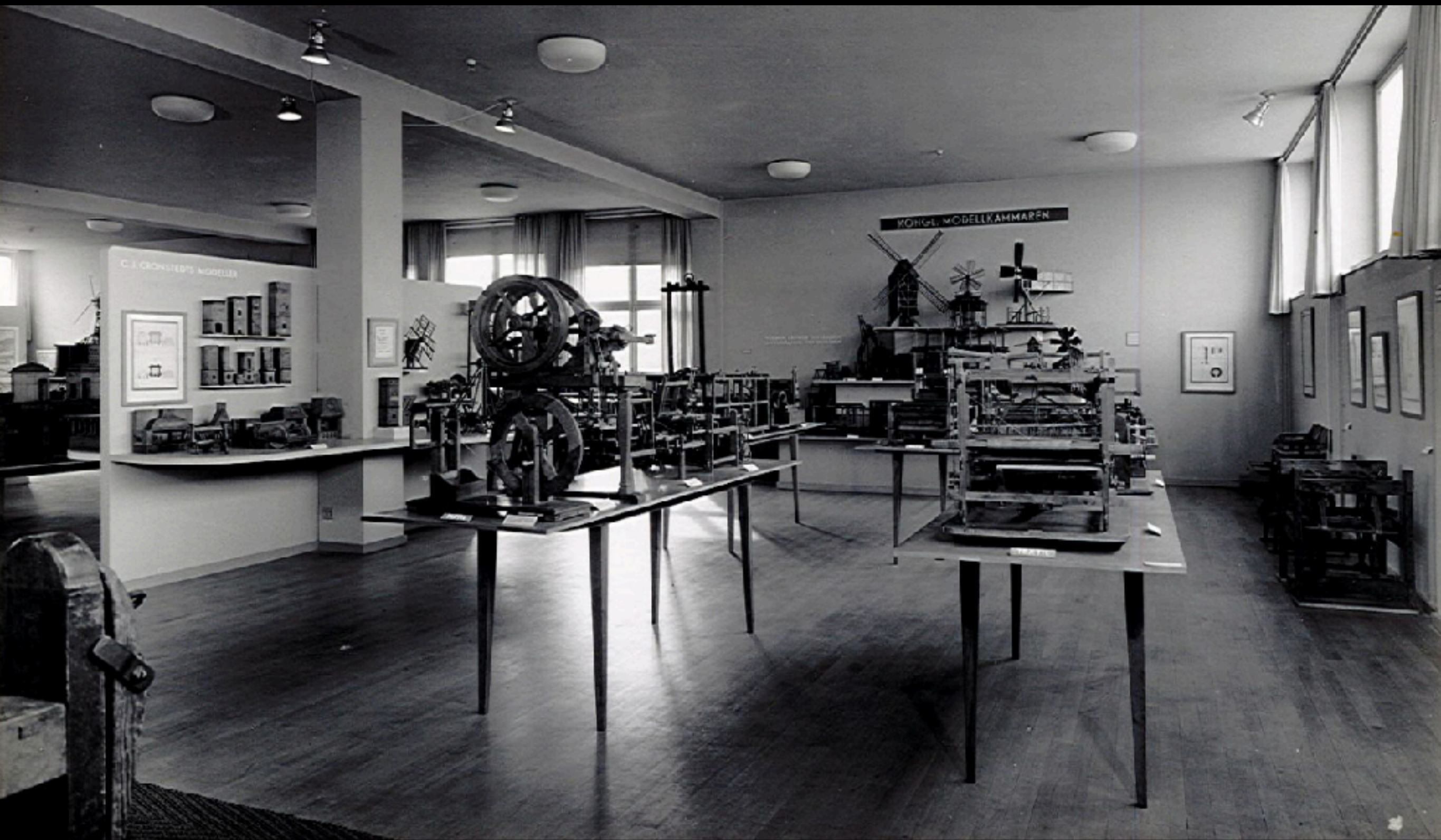
INVENTARIUM  
ÖFVER DE  
MACHINER  
OCH  
MODELLER,  
SOM FINNAS  
VID  
KONGL. MODELL-KAMMAREN  
I STOCKHOLM,  
BELÅGEN UTI GAMLA KONGSHUSET  
PÅ  
K. RIDDAREHOLMEN.



STOCKHOLM,  
TRYCKT HOS ANDERS J. NORDSTRÖM. 1779.

Book frontispieces of Carl Knutberg's, *Tal om nyttan af et laboratorium mechanicum, hållit för kongl. vetenskaps akademien* (Stockholm, 1754), as well as the inventory (of models and machines) at the Royal Swedish Model Chamber in 1779 (compiled by Jonas Nordberg), *Inventarium öfver de machiner och modeller, som finnas vid kongl. modell-kammaren i Stockholm, belägen uti gamla kongshuset på k. Riddareholmen* (Stockholm, 1779).





A representation of the Royal Swedish Model Chamber (with some original models) displayed at the Swedish National Museum of Science and Technology in 1947.





1. Introduction

2. From Archival to Data Driven Humanistic Research

3. About the Research Project "Digital Models"

**4. Textual Models of the Past**

5. Visual Models of the Past

6. Conclusion

All slides in the form of a PDF can be found at <http://pellesnickars.se/>



The **meta-museological aspect** of our project is as stated accentuated by the digitisation of the more than 80 museum yearsbooks—*Daedalus*, published between 1931 and 2015—resulting in **a minor cultural data set** with some 15,000 pages and a few million words.





## Digitala modeller

Introduktion

Daedalus 1931-2015

Sök i Daedalus

Daedalus som dataset

# Daedalus 1964

Meny ☰

In English

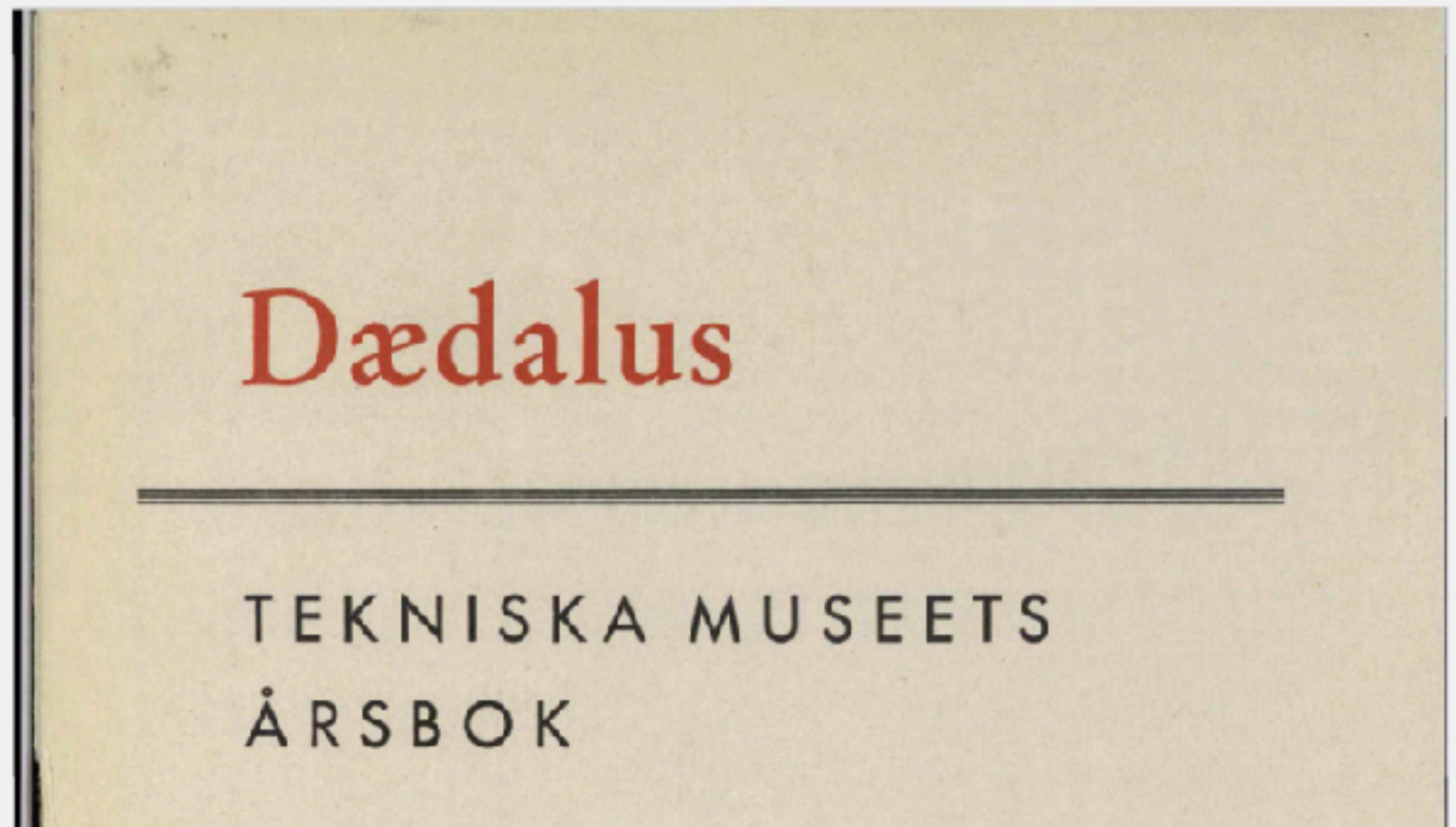
## Artiklar

📄 Ladda ner [pdf](#) | [xml](#) | [txt](#)

1. Christopher Polhems modell till ett skeppsupphalningsverk
2. Wilhelm Teodor Unge
3. Den optiska telegrafen i Furusund

## Läs online

📄 Ladda ner [pdf](#)





# DAEDALUS

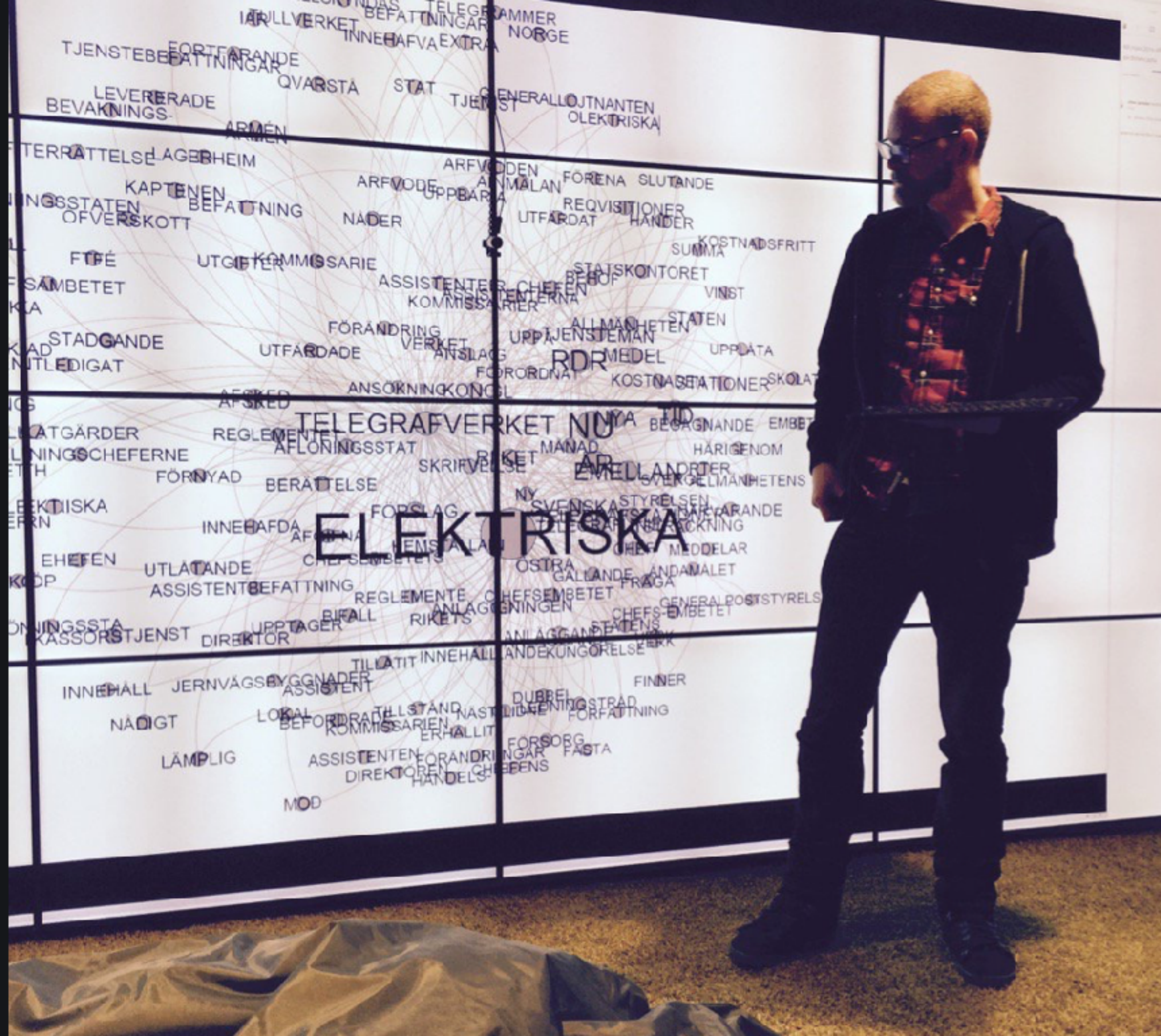
# 1966

Tekniska Museets Årsbok



The plan is to use different types of text analyses tools and methodologies to work and wrangle this data—i.e **algorithmic models and packages** such as **topic modeling, NER-analyses** and **word2vec**.





ELEKTRISKA



na gräs och Silfpackersgräs, samt vid bäret af Högherps- och Tullports-  
området, bäret af Gamla Knäpöskeshuset; i C. M. A. Thomsens  
Gård, 20 St vid Storgatan; i P. M. Hults vid Nybergsgatan å Lads-  
båriga glänsen sättes äppens inflores. Under 10 p i barken inflores

EXPEDITIONEN till Italien, Roms belägring och besättning lara saker som en gång skett och ej kunna stras om. Från de katolska intressens synpunkt och med afseende på den katolska kristenhetens regering är det nödvändigt att påfven återställa på sin fromt

UPPSALA  
C. v. LIN

UFSALA  
r vid Mölles  
kon förhålling

Resultat för att hålla sig kvar i Rom, så länge som en annan makt bibehåller sig i Romagnen.

Amman! Årger! Bland de olika anslagen fanns äfven en proposition angående 1,415,630 fr. till beständande af de utgifter som erfordras för okkupationsarméns underhåll i Italien under de sex näst påseendena af

Härnå gång som analog begäres för expeditionen till Rom är det en strängt menande pligt för oss att åstadga trösten: Hjärtspjettande och förförande vår protest mot själva expeditionens princip, höra vi prövas och skänkas för både kända och okända.

...som stundfärdiga den som vi föraktat, som återkom  
fjerdans regeringen. (Röst på vänstra sidan. Teo-  
nen till ställighet på den högra.)

Fruenslingen står på nära slutet af sina arbeten  
före serien. Jag kan förstå att den ej är benägen  
för långa diskussioner. Min mening är heller inte

*Råter från bögra sidan: NÄ, EN saken då! Tala.  
Ersattel åttan-Den badsmåda införandet, som en*

dogget för skottets öfugt i ämnet, är general Gros-  
chy, ytttrar i sitt betänkande:

»Från Frankrikes syngare, med afsende på dess  
skeddsvärdens följda politik och dess behof att upp-  
rätthålla sin befästade och välskötta öfver national-

stärker  
stärker  
stärker

Franskrikes inflytande, mina herrar — den fransk  
regeringens inflytande i Italien är — intet! (I Knot p  
1898 sidan 1). Det är skändligt!

*Emmanuel Arago berättar:*  
Och då jag bekräftar detta, så är det icke med syn

... Sundev II, ...  
... HG, ...  
... General ...

der, huru man i korn skäpar lag och rätt och hur  
andets bästa medborgare der äro ständiga föremål  
för förföljelser och grymhet. (Reklamationer u

OTHLAND, Jag vill ej ånyo skildra detta olyckliga tillstånd.  
En omständighet torde likväl fortjona att framdragas

ning af öfver hela Europa och ådagalagt att det härskande inflytandet i Rom är Österrikes och Neapel kändesvald. (Svenskan.)

...kænnen Hans Helligbets rest til Castel Gandolfo  
riedberg, 6 b i viden att han der träffade kungen af Nicapel,  
hvaras af en Österrikisk general. När man ser på

**GSOI R**

... tillfredt gör talsaren en exposé af Gladstones be-  
 stanta bevis till loed Aberdeen om alla de barbariska  
 brottligheter, som i Nepal beris mot män, slöso

ART. — Oh, oh! — Ennn. Arago fortæller:

(Horse shoe.)



# Cultural heritage as digital noise: nineteenth century newspapers in the digital archive

Johan Jarlbrink and Pelle Snickars

*Department of Culture and Media Studies, Umeå University, Umeå, Sweden*

1228

Received 5 September 2016

Revised 2 March 2017

Accepted 19 March 2017

## Abstract

**Purpose** – The purpose of this paper is to explore and analyze the digitized newspaper collection at the National Library of Sweden, focusing on cultural heritage as digital noise. In what specific ways are newspapers transformed in the digitization process? If the digitized document is not the same as the source document – is it still a historical record, or is it transformed into something else?

**Design/methodology/approach** – The authors have analyzed the XML files from *Aftonbladet* 1830 to 1862. The most frequent newspaper words not matching a high-quality references corpus were selected to zoom in on the noisiest part of the paper. The variety of the interpretations generated by optical character recognition (OCR) was examined, as well as texts generated by auto-segmentation. The authors have made a limited ethnographic study of the digitization process.

**Findings** – The research shows that the digital collection of *Aftonbladet* contains extreme amounts of noise: millions of misinterpreted words generated by OCR, and millions of texts re-edited by the auto-segmentation tool. How the tools work is mostly unknown to the staff involved in the digitization process? Sticking to any idea of a provenance chain is hence impossible, since many steps have been outsourced to unknown factors affecting the source document.

**Originality/value** – The detail examination of digitally transformed newspapers is valuable to scholars depending on newspaper databases in their research. The paper also highlights the fact that libraries outsourcing digitization processes run the risk of losing control over the quality of their collections.

**Keywords** Sweden, Archives, Documents, Accuracy, Print media, Auto-segmentation, Character recognition equipment, Large-scale digitization

**Paper type** Research paper

## Introduction

In October 1847, the telegraphic wire in St Germain outside of Paris was struck by lightning. The Swedish newspaper *Aftonbladet* reported that a telegraph assistant at a station nearby had discovered the demolished telegraph printing several letters on its own. Yet, according to the paper, “since they were not coherent, he decided to signal the phrase used for ‘I do not understand.’” In doing so, however, he received “a heavy electric shock, which was followed by a loud bang, sounding like a gunshot” (*Aftonbladet*, 1847).

Within a digitization project initiated by the National Library of Sweden, the 1847 October copy of *Aftonbladet* was digitized in 2013 at the Swedish Media Conversion Centre. The newspaper *Aftonbladet*, founded in 1830, was one of the key titles in nineteenth century Sweden. It is often described as the first modern newspaper – consequently, it was also the first newspaper to be completely digitized by the National Library. Then again, if a telegraph struck by lightning in the late 1840s produced some real uncanny results, the same can be said of present day digitization processes. The digital version of the paper with the lightning-telegraph incident, in fact, literally reported that the struck assistant “saw a dazzling light along the wires on the walls conducting electricity de visl devärdigavard värdigavard dejennte fullkomen ihåfvintparkerslagna förvintparkerslagna parkerslagna ken – tas till 70 70 misvärt fruktarsnart tAf och sisrans njes ej [...] which fell down in pieces, burning the table and the floor.”

This research has been funded by The Torsten Söderberg Foundation.





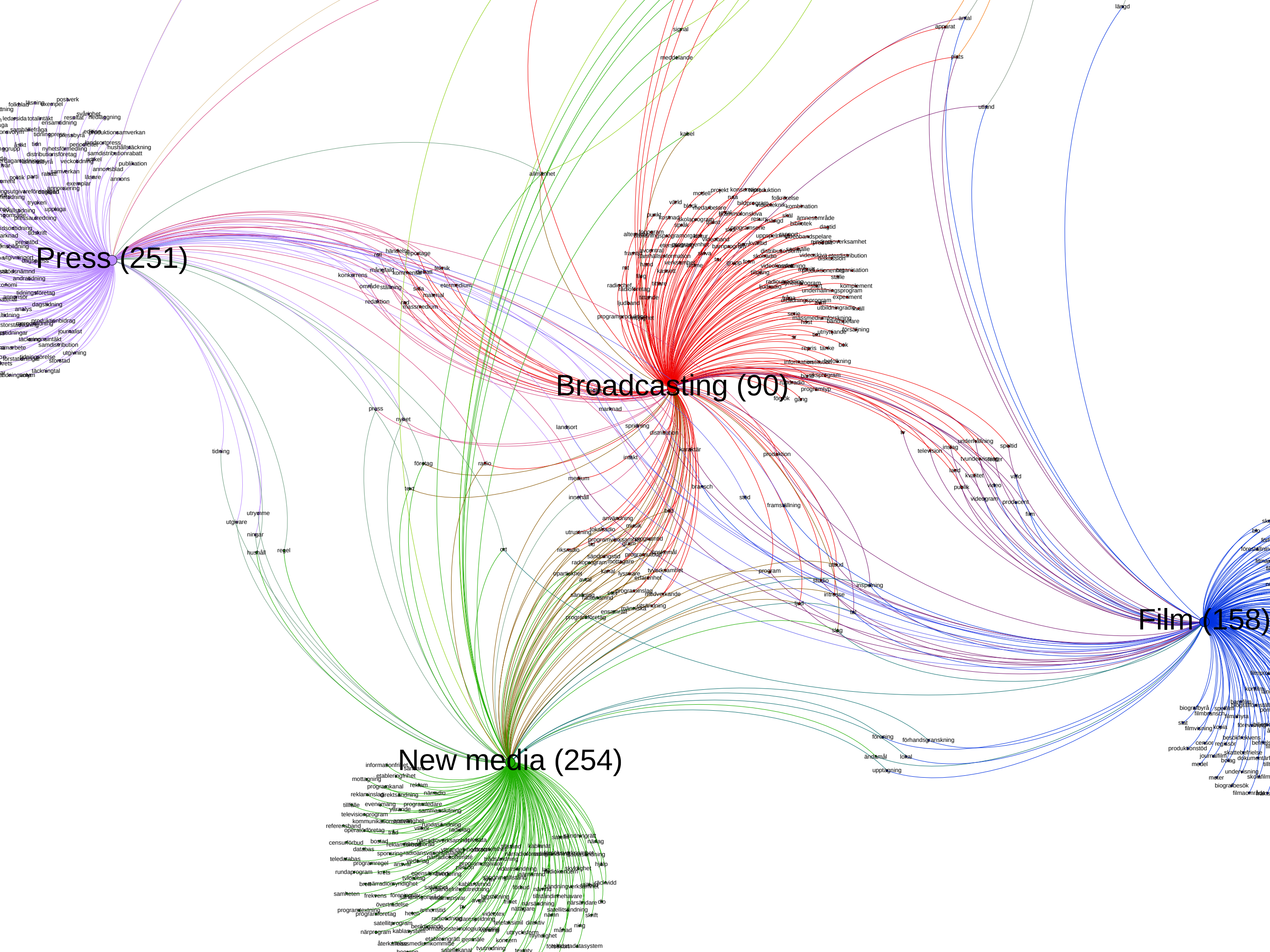




# Film (158)

granskningsverksamhet kopiering djur handling biografägarförbundet medarbetare  
komponent biobesökare anspelning hemvideogram  
rekommendation journal medlem reportage pliktexemplar  
nöjesskattutredningen filmlaboratorium kassett programråd nöjesfilm filmforsknings biografnäring  
visningsnämnd producentandel bemärkelse filmvetenskap filmföretag premiärbiograf  
produktionsledare matiné tillställning publikfrekvens filmföretag premiärbiograf  
filmproducentförening visningsmöjlighet reklam åldersgrupp gräns filmföretag premiärbiograf  
videomarknad granskare barnmatiné våldsfilm filmföretag premiärbiograf  
kontrollförening granskningsman lustspel plats nöjestillställning tv-undervisning filmföretag premiärbiograf  
uppfostningssyfte filmkamera filmföretag premiärbiograf  
rekvisita kamera speltid rådgivning filmföretag premiärbiograf  
biografnöjesskatt censurgrund filmföretag premiärbiograf  
efterhandsstöd gård format filmfond besöksfrekvens våldsskildring långfilm biografägare filmvisning inslag medium föreförande  
fantasi organ premiär regissör filmmakiv våldsskildring långfilm biografägare filmvisning inslag medium föreförande  
verksamhetsberättelse produktionsföretag svenska färgfilm spelår stat produktion nöjeskatt barnfilm filmproduktion filmverksamhet verklighet biografverksamhet restitution språk  
filmcentrum graf våldinslag ateljé antal land  
cinematek visningkopia ort bolag filmbolag stöd  
instruktionsfilm konst godkännande biografägareförbundet författare  
prestation händelse betänkande skola  
filmrepertoar filmarbete befrielse intäkt  
inköp filmpåverkan tillträde import  
visningsställe filmalag filmcensurutredning publik  
stillbild filmbild upplysning vecka visning  
opinion konstart annonsfilm genre kvalitetsfilm  
premiering proportion folk kvalitet  
förhandsstöd bruttopris filmotek biobesök filmcensur  
cheflön säsong nöjesbeskattning åskådare  
upplevelse filmprogram kategori videogram  
socialfilm förvaltningsråd förevisning  
helhet statsanslag dokumentärfilm  
bildserie förfogande filmstöd förening  
upphetsning skräck granskningssavgift filmproducent  
produktionsgaranti hyra undervisning skildring  
smalfilmsdistribution sällskap undersökningsperiod biograförelse producent film  
filmintresse filmgranskning filmutredning  
enbiografföretag kulturfilm anslag klipp inspelning framställning biografbyrå institut filmbransch filmindustri  
katalogisering underhållningsfilm skådespelare spridning spelfilm biografbyrå institut filmbransch filmindustri  
dokumentär minut ägare reklamfilm bedömning intresse förhandsgranskning produktionskostnad skadeverkning biografbyrå institut filmbransch filmindustri  
arkivering musik avgift bruttointäkt landsort repertoar förhandsgranskning produktionskostnad skadeverkning biografbyrå institut filmbransch filmindustri  
filmprojekt nöjeskattsbefrielse biografbransch filmare biljett journalfilm föreningsbiograf stumfilm bibliotek bruk  
skolungdom duk granskningsråd filmområde television ändamål titel barnfilmskommittén skatteåterbäring visningsverksamhet  
matinéprogram filmstudiorörelse projektionsapparat underhållning biografbesök filmpolitik täckningsbidrag filmkommitté produktionsutläggning el  
delbetänkande filmgranskningråd visningsrätt filmcentral filmproduktion samproduktion filmråd biografväsande uppbördsmyndighet  
spelfilmsproduktion filmnämnd färg barncensur vuxencensur tv produktionsbidrag återbäring barnfilmsklubbar visningstid kontrast filmregissör  
upphovsman barnfilmsproduktion anordnare filminspelning filmorganisation långfilmsproduktion filmklubb excerpt  
filmföreställning filmföreställning förhandsgranskning utrustning branschorganisation filmtyp visningsställe stödsystem  
alster föreningsliv frakt distributionkostnad utbyte videovåld smak tillståndskort utlåning foto  
fyllnadsfilm beställningsfilm utbyte videovåld smak tillståndskort utlåning foto  
visninglokal fraktkostnad censurmyndighet videokassett hjälte bruttobiljettintäkt  
visningsorganisation nöjeskatteåterbäring filmfestival skådespel







15. feature article  
16.

17. **FREDRIK NORÉN AND PELLE SNICKARS**  
18. Umeå University  
19.  
20.

## 21. **Distant reading the history of** 22. **Swedish film politics in 4500** 23. **governmental SOU reports** 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34.

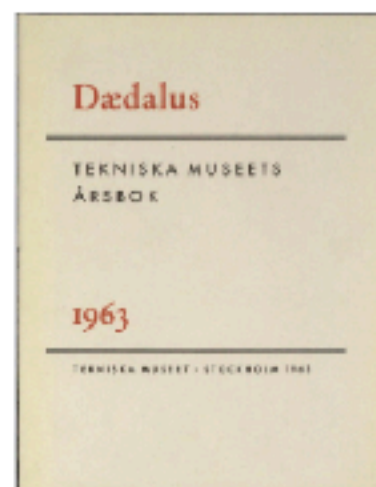
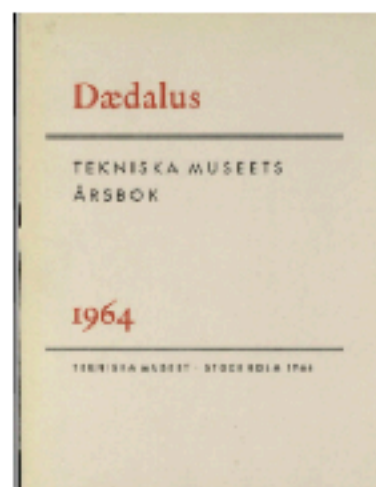
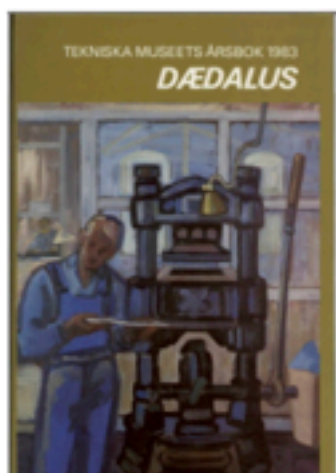
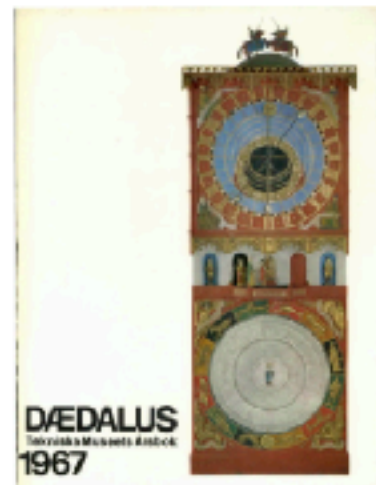
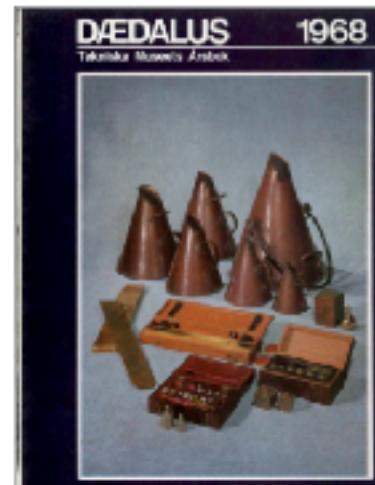
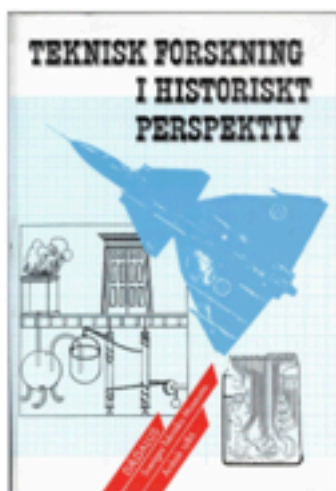
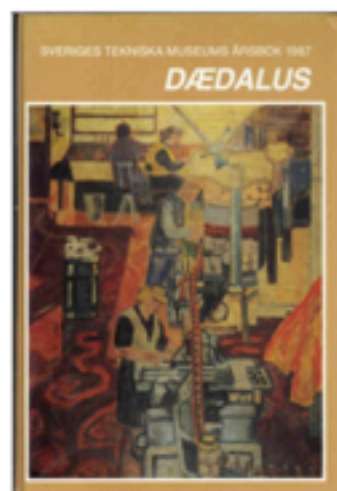
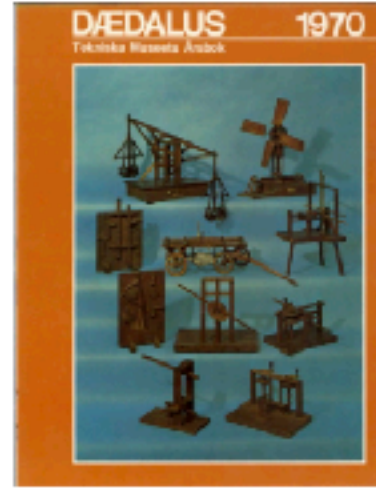
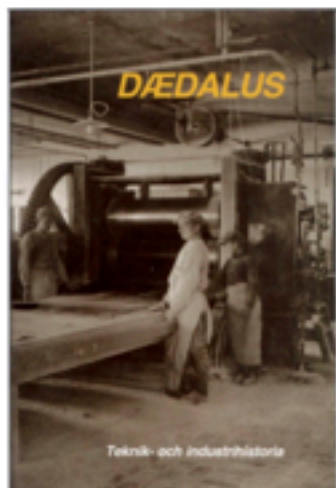
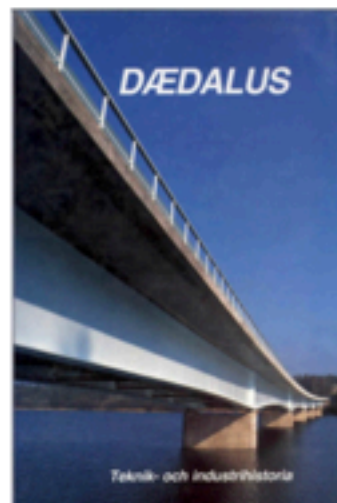
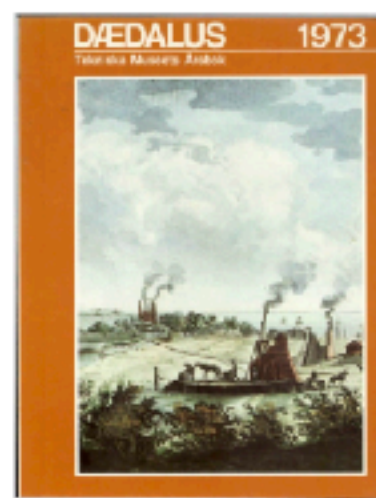
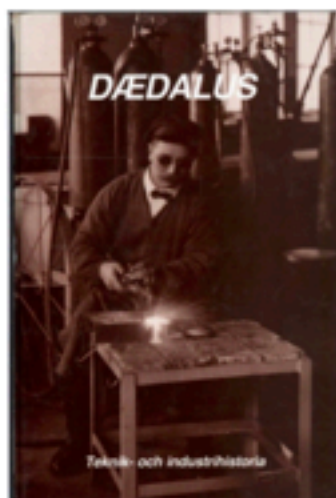
### 35. **ABSTRACT**

36. *Using computational methods, digitized collections and archives can today be*  
37. *scrutinized in their entirety. By distant reading and topic modeling one particu-*  
38. *lar collection – 4500 digitized Swedish Governmental Official Reports (SOU)*  
39. *from 1922 to 1991 – this article gives a new archival perspective of the history of*  
40. *Swedish film politics and policy-making. We examine different probabilistic topics*  
41. *related to film (and media) that the algorithm within the topic modeling software*  
42. *Mallet extracted from the immense text corpora of all these Official Reports. The*  
43. *article methodologically recounts and analyses novel ways to understand the*  
44. *history of Swedish film politics and policy-making through topic modeling the*  
45. *massive text corpora of SOU publications. Topic modeling is a computational*  
46. *method to study themes in texts by accentuating words that tend to co-occur and*  
47. *together create different topics. Basically, it is a research tool for the discovery of*  
48. *hidden semantic structures, exploring a collection through the underlying topics*  
49. *that run through it. Hence, our article captures a number of film discourses and*  
50. *trends within the SOU material. In conclusion, we argue that topic modeling*  
51. *should be recognized as a method and research aid for gathering an overview of a*  
52. *major material; as a way to pose new and unforeseen research questions; and as a*

### **KEYWORDS**

digital methods  
digitized collections  
and archives  
history of Swedish film  
politics  
distant reading  
topic modeling  
Swedish Governmental  
Official Reports  
(SOU)  
computational film  
studies







## Textual Models of the Past

The industrialisation of Sweden put brand **new geographical places on the map**. Towns grew up around mines, saw mills, and mechanical factories all over the country.

Yet, precisely **where is the national history of technology located geographically** when presented by Daedalus? And how does this **historiographical geography change over time**?

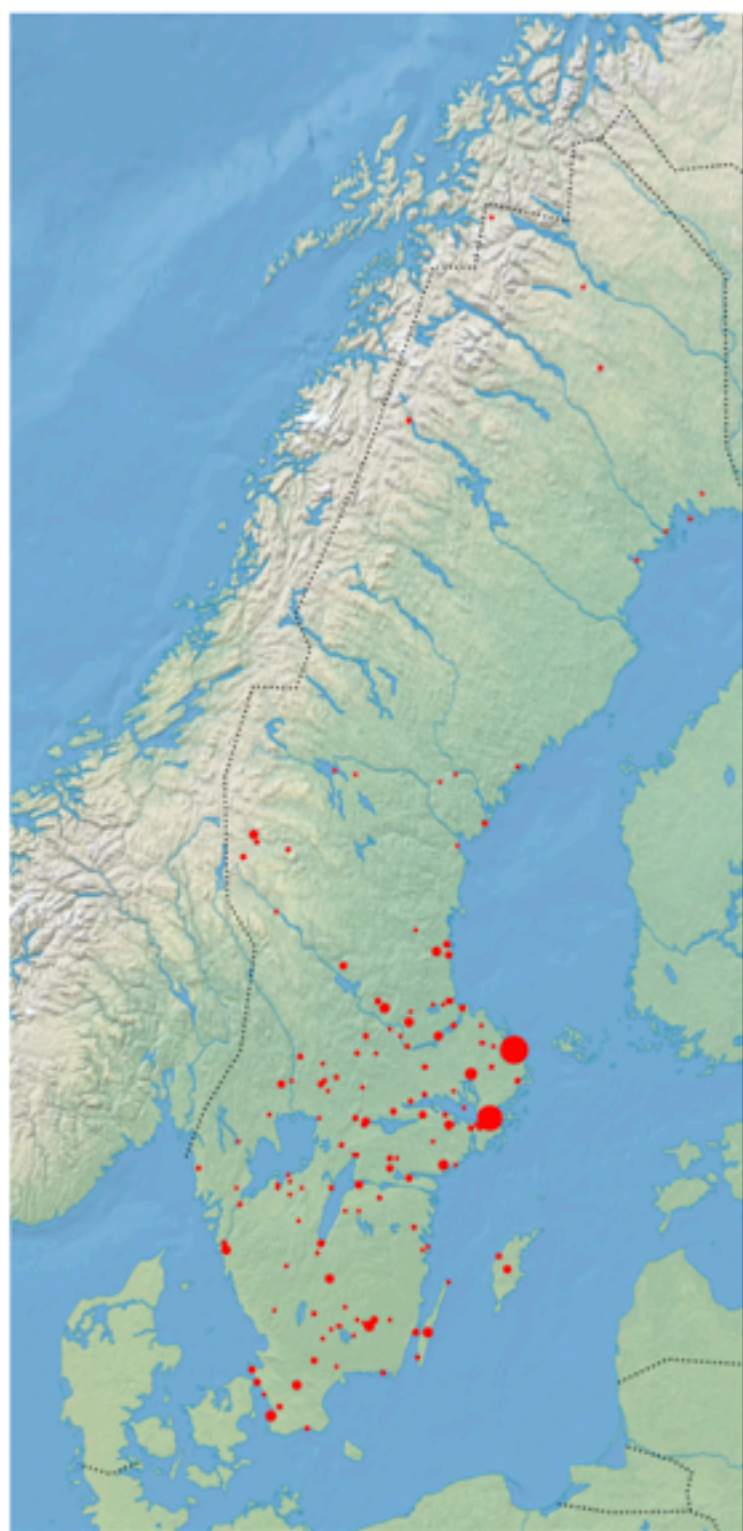


## Textual Models of the Past

We extracted the texts from the alto-xml files and cleaned the texts manually in order to get rid of place names in bibliographies, advertisements etcetera.

A **named-entity recognition (NER) implementation** for Swedish was used to extract place names from the cleaned texts. Unique hits were checked, false hits deleted and the place names assigned GIS coordinates (using Google Maps Geocoding API).





1930s



1980s



2010s

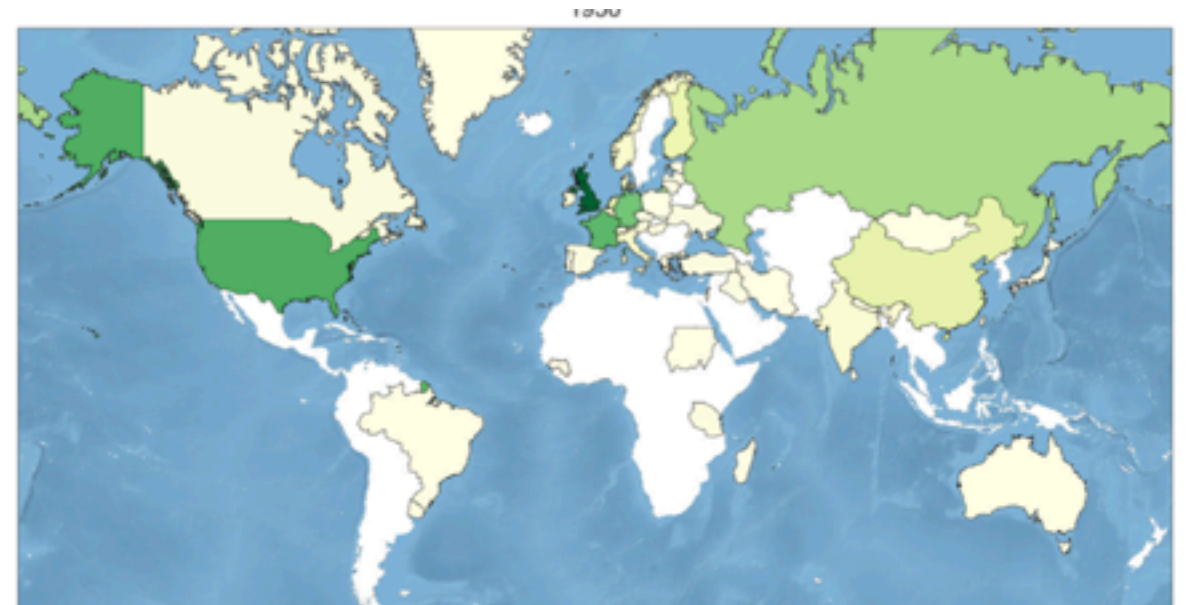


Textual Models of the Past—

## Where is the World of Technology (according to Daedalus)?



1930s

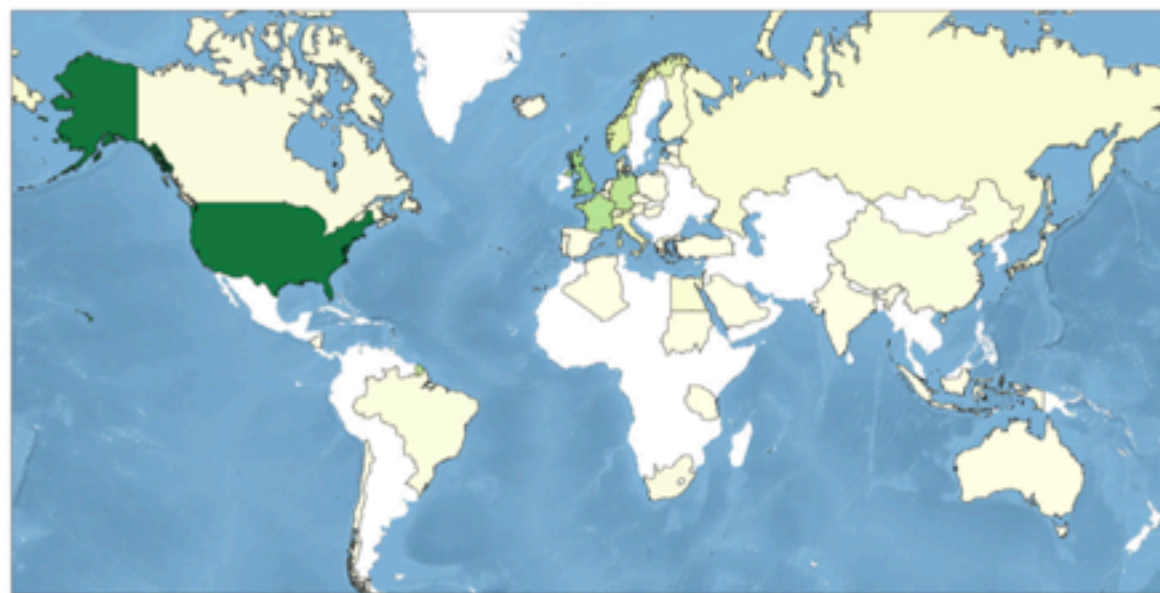


1950s

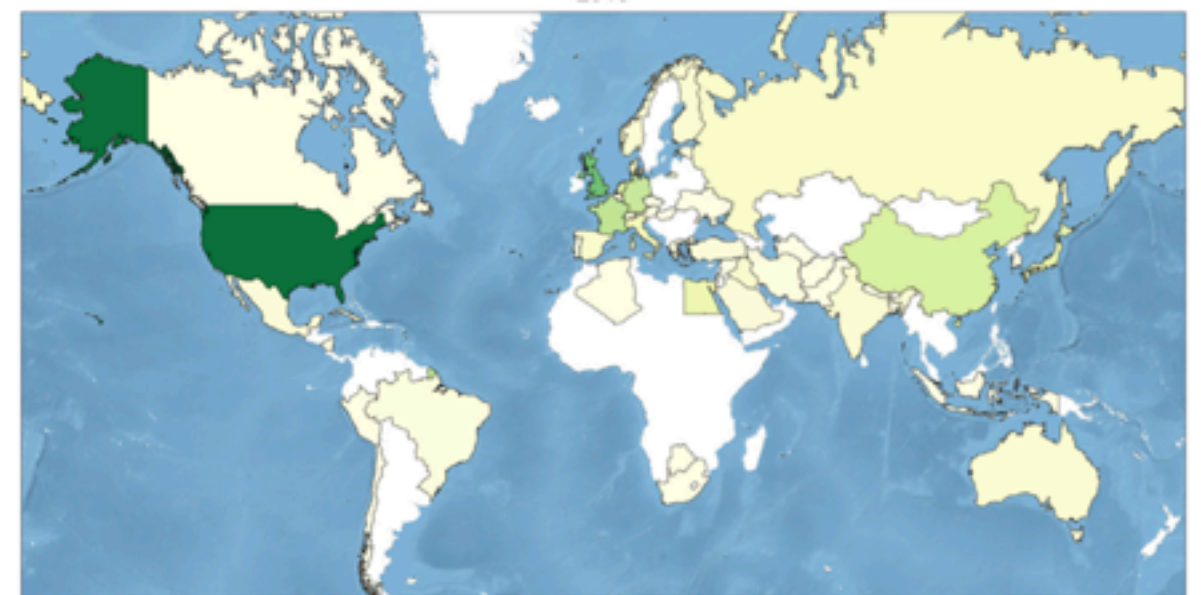


Textual Models of the Past—

**Where is the World of Technology (according to Daedalus)?**



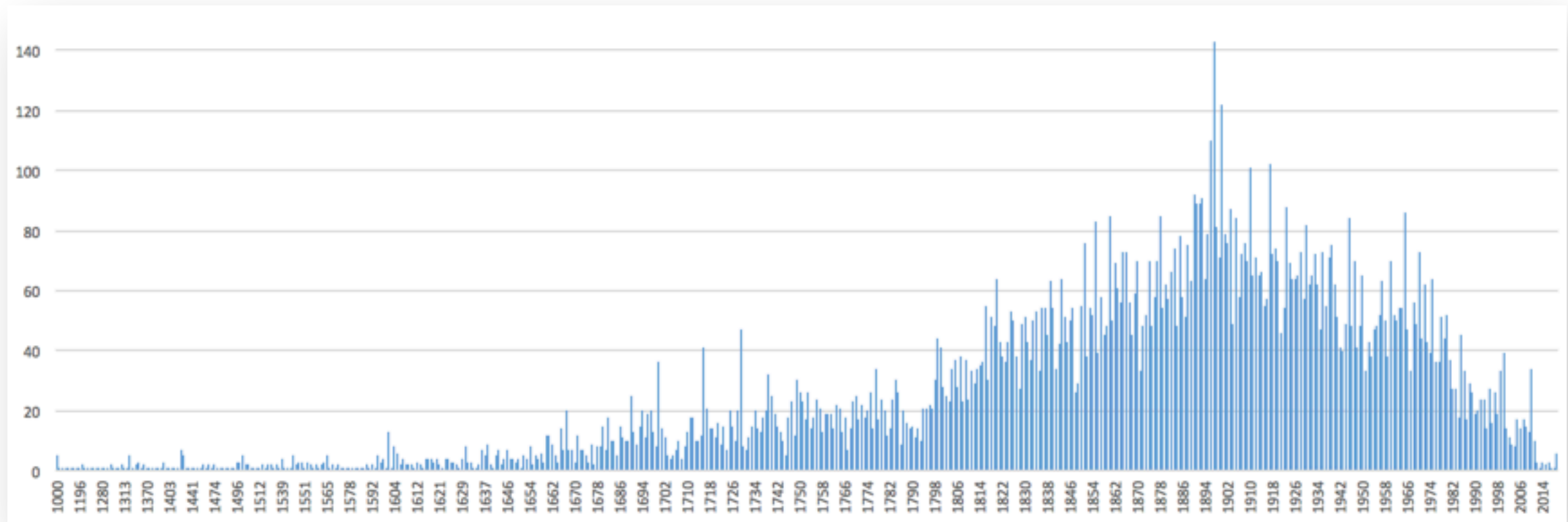
1980s



2010s

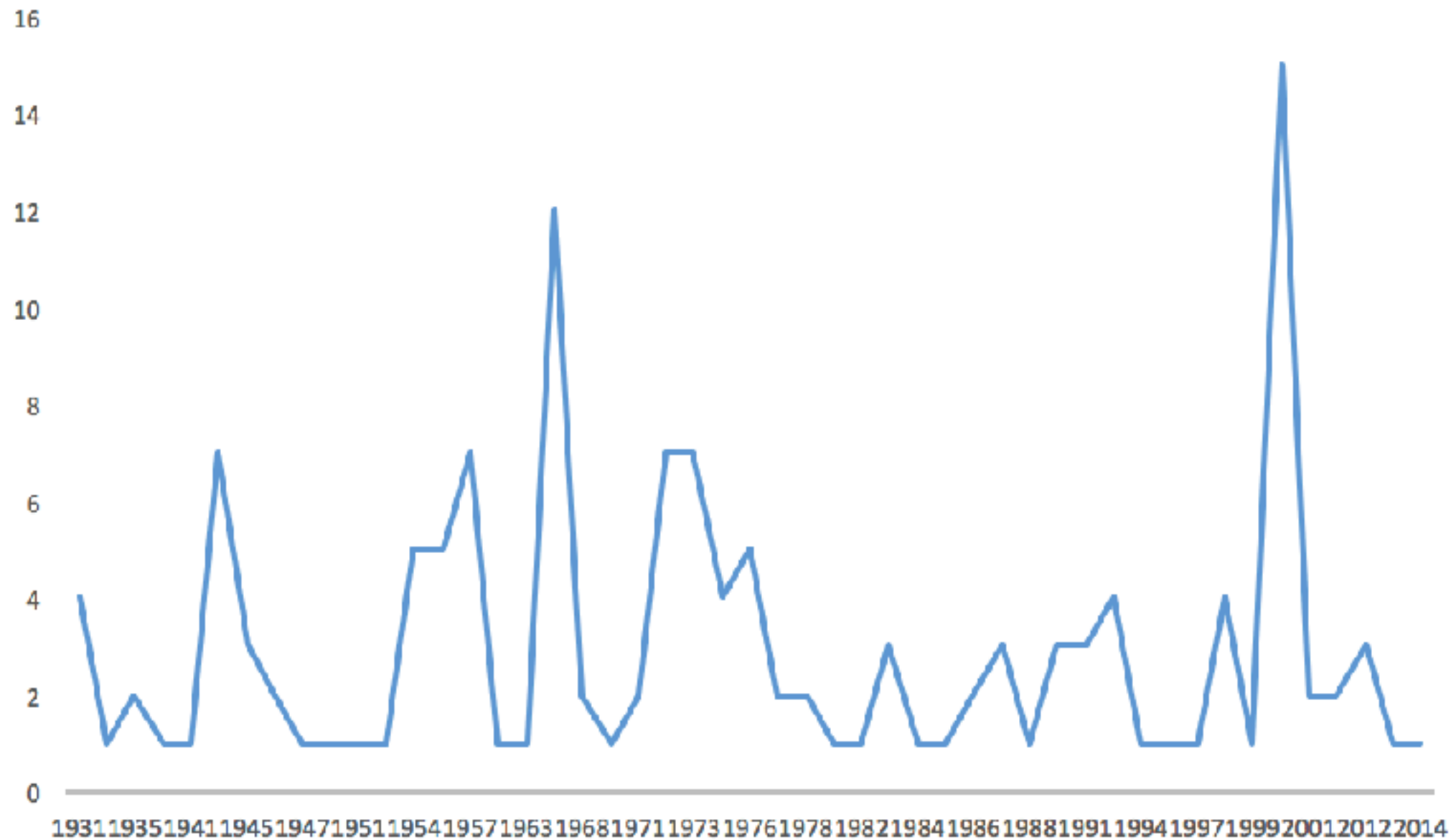


**When is the history of technology?**—most frequent years mentioned in Daedalus.



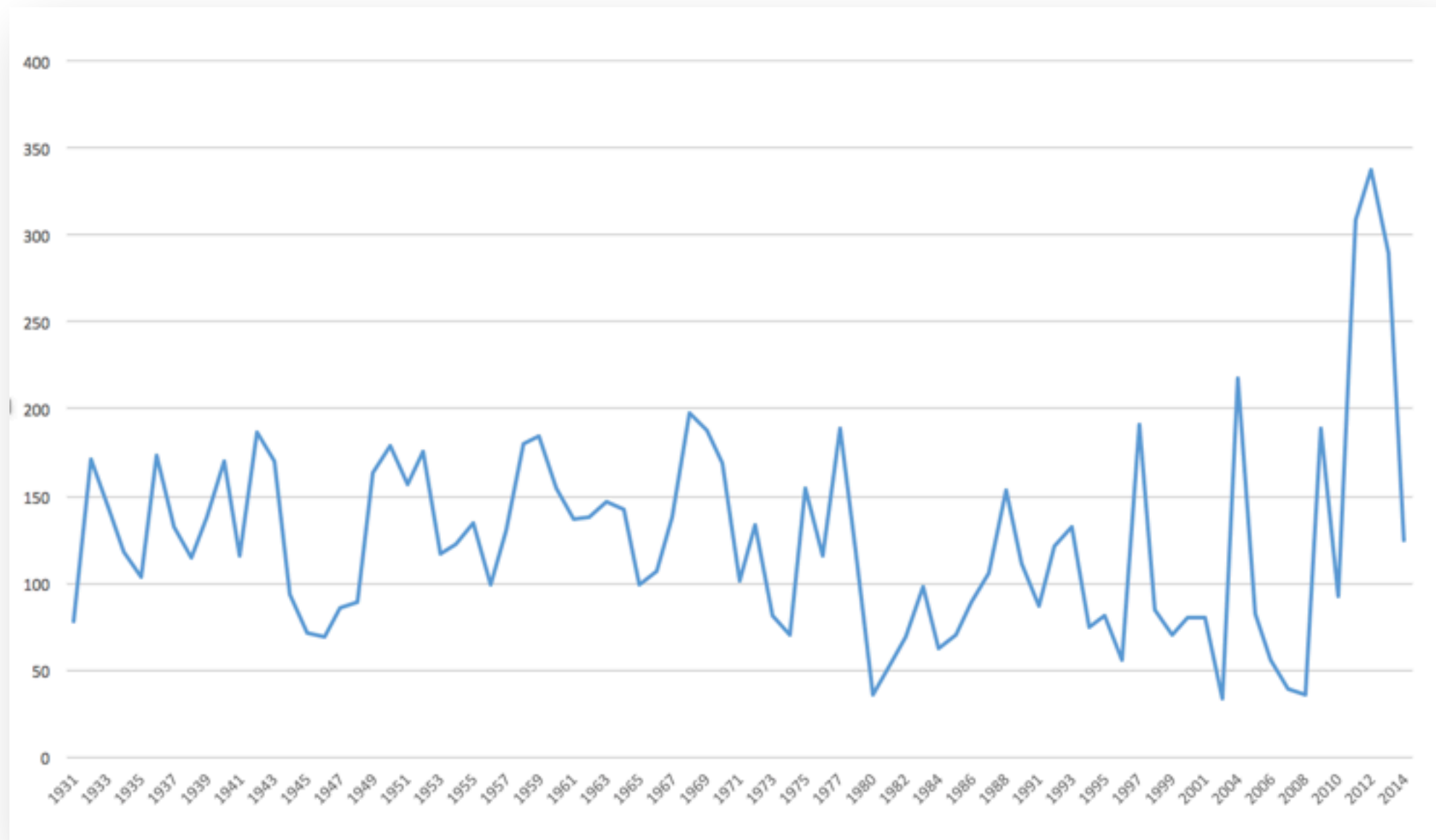


How frequent are years mentioned in Daedalus—as for example 1897?





And what about the **time horizon of the history of technology** (according to Daedalus)? Amount of years that each yearbook looks back in time.





## Textual Models of the Past

At present we are working with preparing the Daedalus data set for **topic modeling**, where articles need to be clearly segmented.

We have also begun—in collaboration with our programmers at Humlab—to try to understand if and how **word-embedding models** can be useful.



## Textual Models of the Past

At present we are working with preparing the Daedalus data set for **topic modeling**, where articles need to be clearly segmented.

We have also begun—in collaboration with our programmers at Humlab—to try to understand if and how **word-embedding models** can be useful.

Some of our first try outs, for example, show that the word vector "**technology**" has an analogy an affinity to "**iron handling**"—which is hardly suprising. But also to a word vector as "**possibilities**", hinting at the ways in which **technology and future are linked concepts**.



The background is a historical manuscript page, likely from a technical or scientific work. It features several detailed drawings of mechanical components. At the top, there is a title in Swedish: "Den första Machinen om Wallnuths Kraft i ställige fall." (The first machine about the power of the walnut in a standing position). Below the title, there are three small wheels labeled 1, 2, and 3. To their right are several other mechanical parts, including a cross-shaped component and a long rod. On the far right, there is a drawing of a frame structure. The bottom half of the page is dominated by a large, complex drawing of a machine, possibly a press or a mill, with a large wheel and a frame. The text "1. Introduction" is overlaid on the left side of the page.

1. Introduction

2. From Archival to Data Driven Humanistic Research

3. About the Research Project "Digital Models"

4. Textual Models of the Past

**5. Visual Models of the Past**

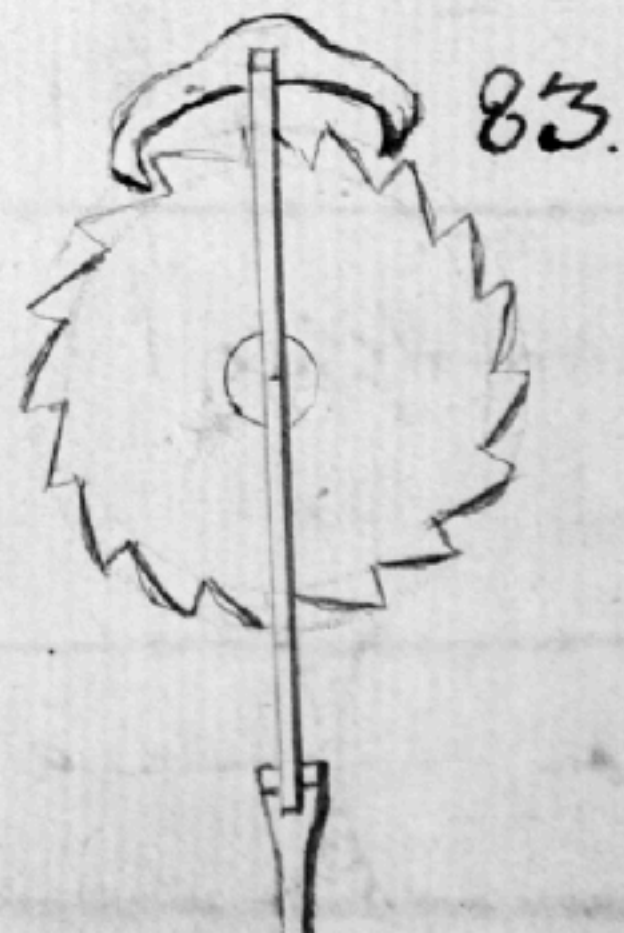
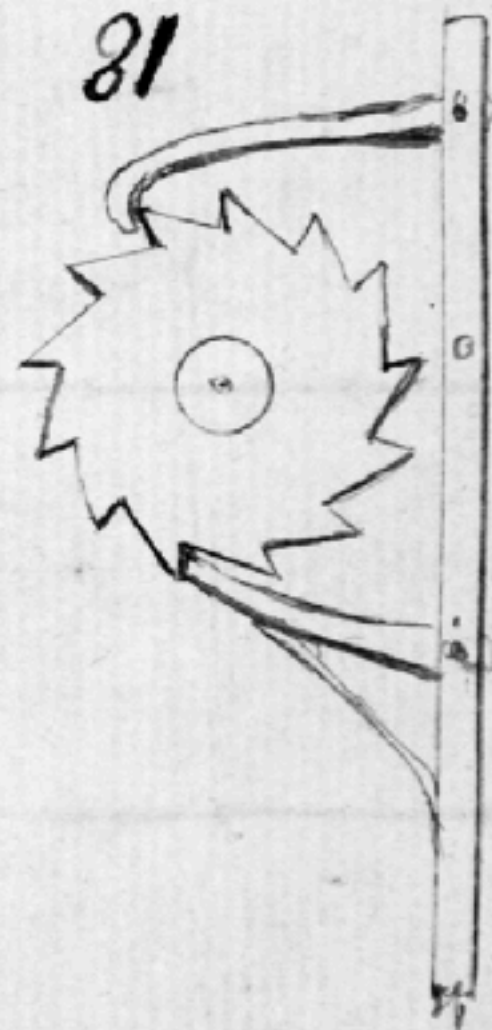
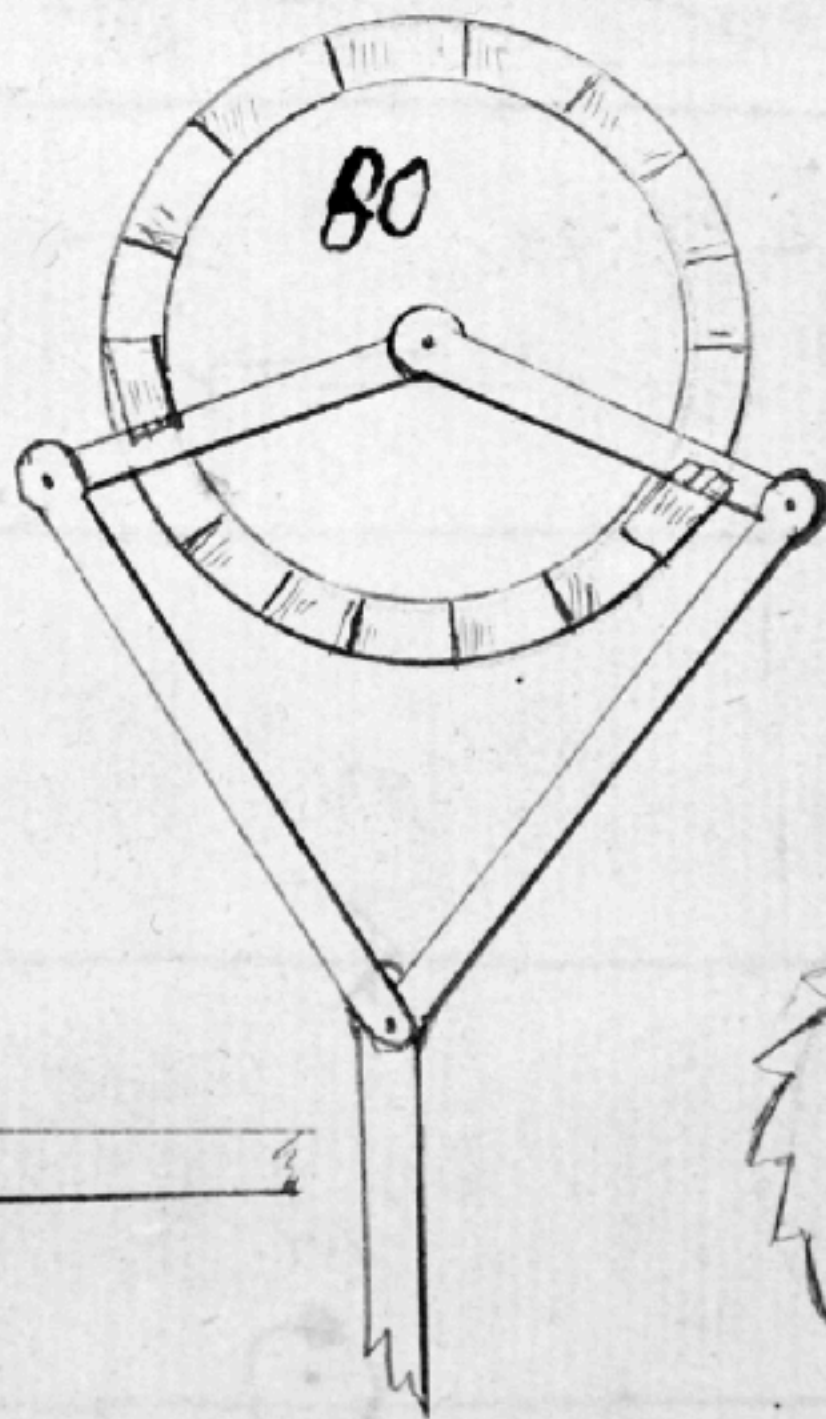
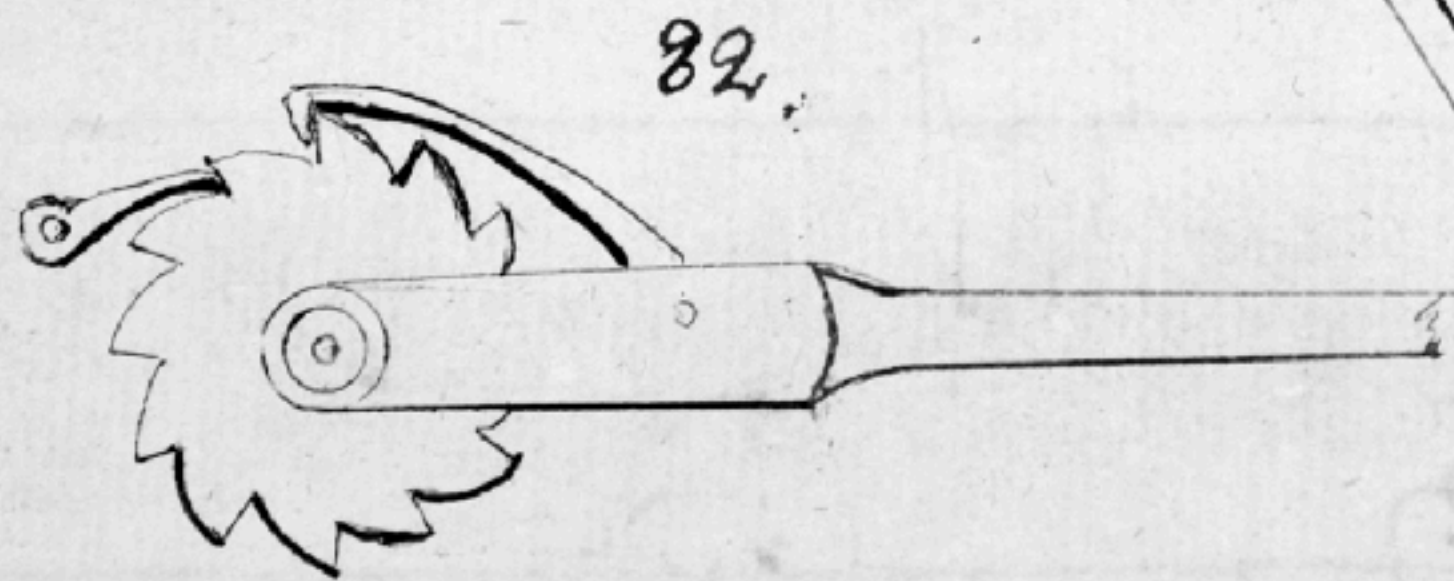
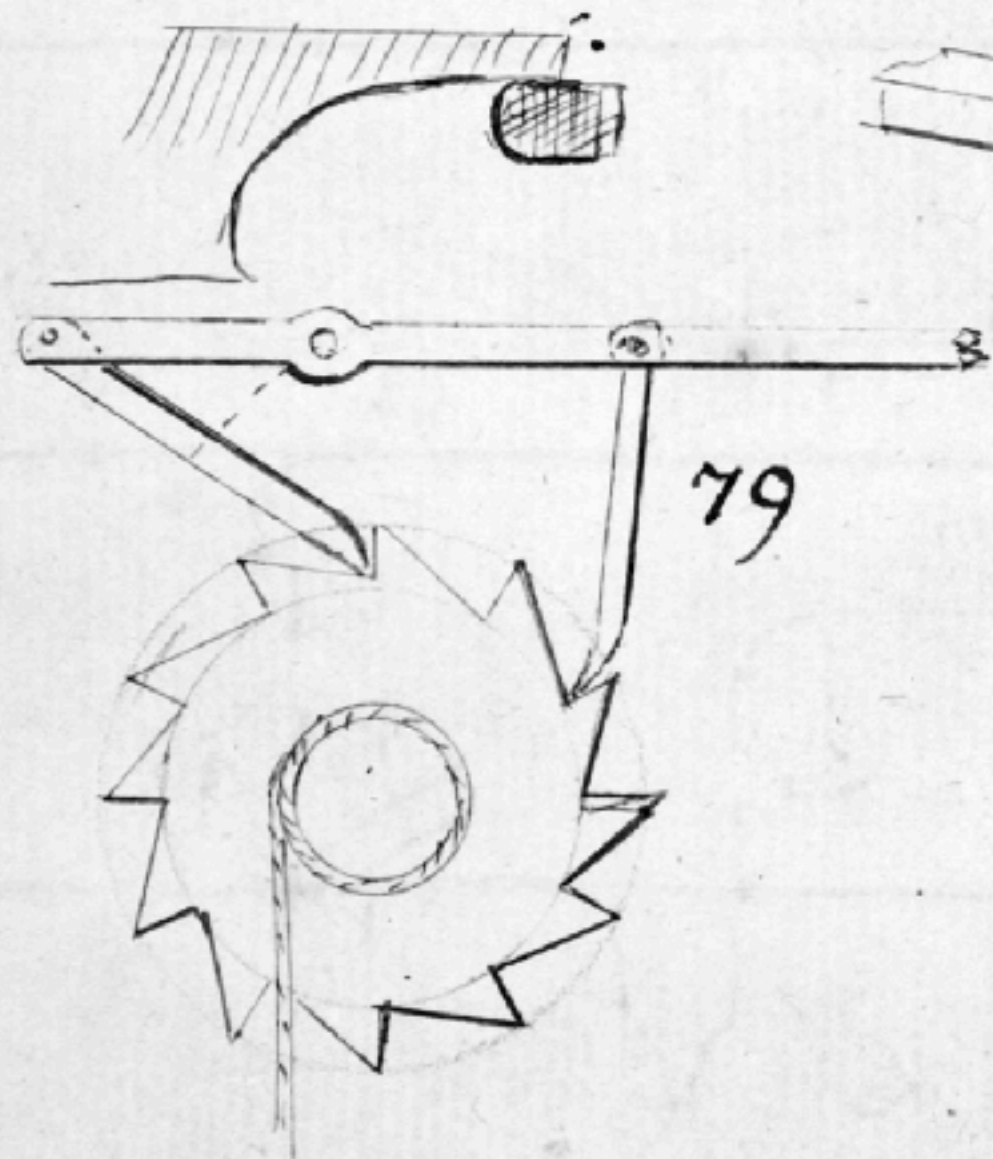
6. Conclusion

All slides in the form of a PDF can be found at <http://pellesnickars.se/>

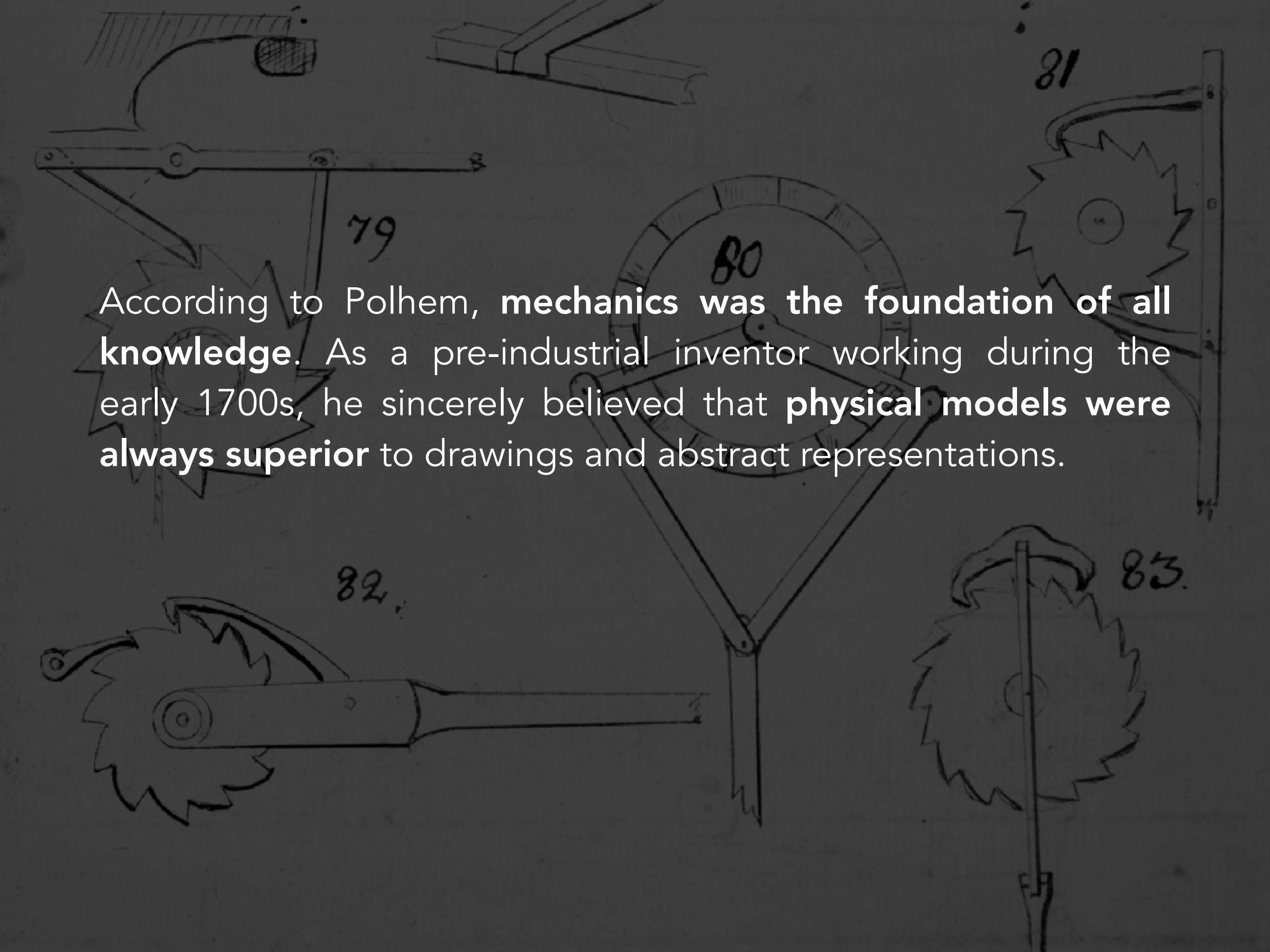










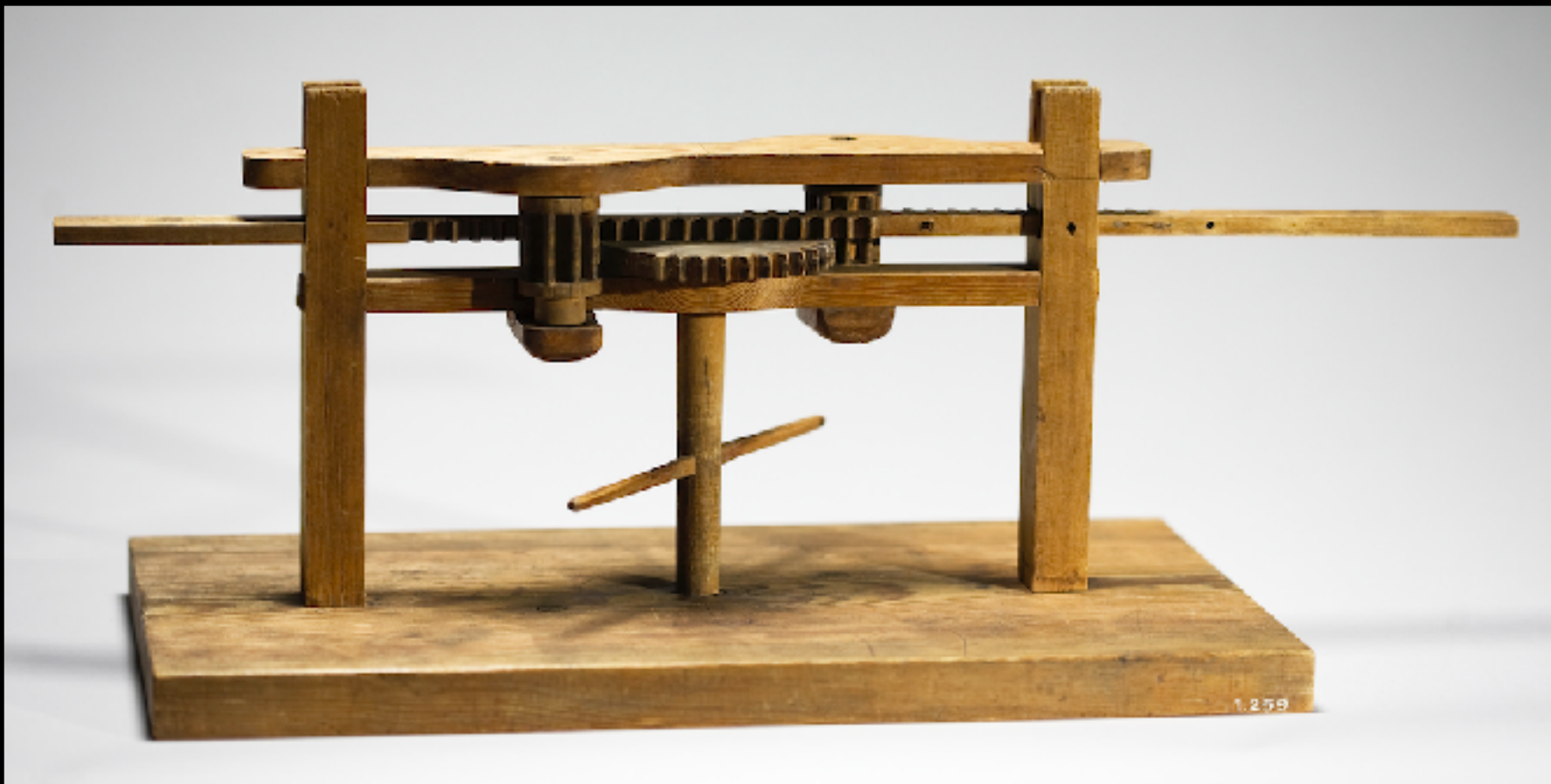
The background of the slide features several hand-drawn sketches of mechanical devices, likely from a historical manuscript. These include: a horizontal beam with a pulley system (labeled 79), a circular dial or clock face (labeled 80), a gear mechanism (labeled 81), a large gear with a long handle (labeled 82), and another gear mechanism (labeled 83). The sketches are rendered in a simple, line-drawn style.

According to Polhem, **mechanics was the foundation of all knowledge**. As a pre-industrial inventor working during the early 1700s, he sincerely believed that **physical models were always superior** to drawings and abstract representations.



Since a writer naturally had to know the alphabet in order to create words and sentences, Polhem argued that a contemporary mechanicus had to grasp **a similar mechanical alphabet** to be able to construct and understand machines. This seems to have been Polhem's main idea for constructing and establishing the different wooden models in his so called **mechanical alphabet**.





Models from Christopher Polhem's "mechanical alphabet" from the early 1700s. Actual models—whether in the form of originals or copies—can today be found at the Mining Museum in Falun as well as at the Swedish National Museum of Science and Technology in Stockholm.







During the latter half of the 18th century these models—and many others—were displayed at **the Royal Swedish Model Chamber**—located at the fashionable Wrangels Palace on Riddarholmen in Stockholm. The institution was open to the public, and counted as one of the **finest physical model collections in Europe**.



## Visual Models of the Past

In order to investigate **the specificity of three-dimensional scanning, rendering and modeling**, we decided **to apply five different forms of 3D visualisations** of Polhem's alphabet—executed in **altered media modalities**.



## I. Stupid Scanning

We used an ordinary iPhone—and the Agisoft Photoscan software—to repeatedly photograph one of Polhem's models.







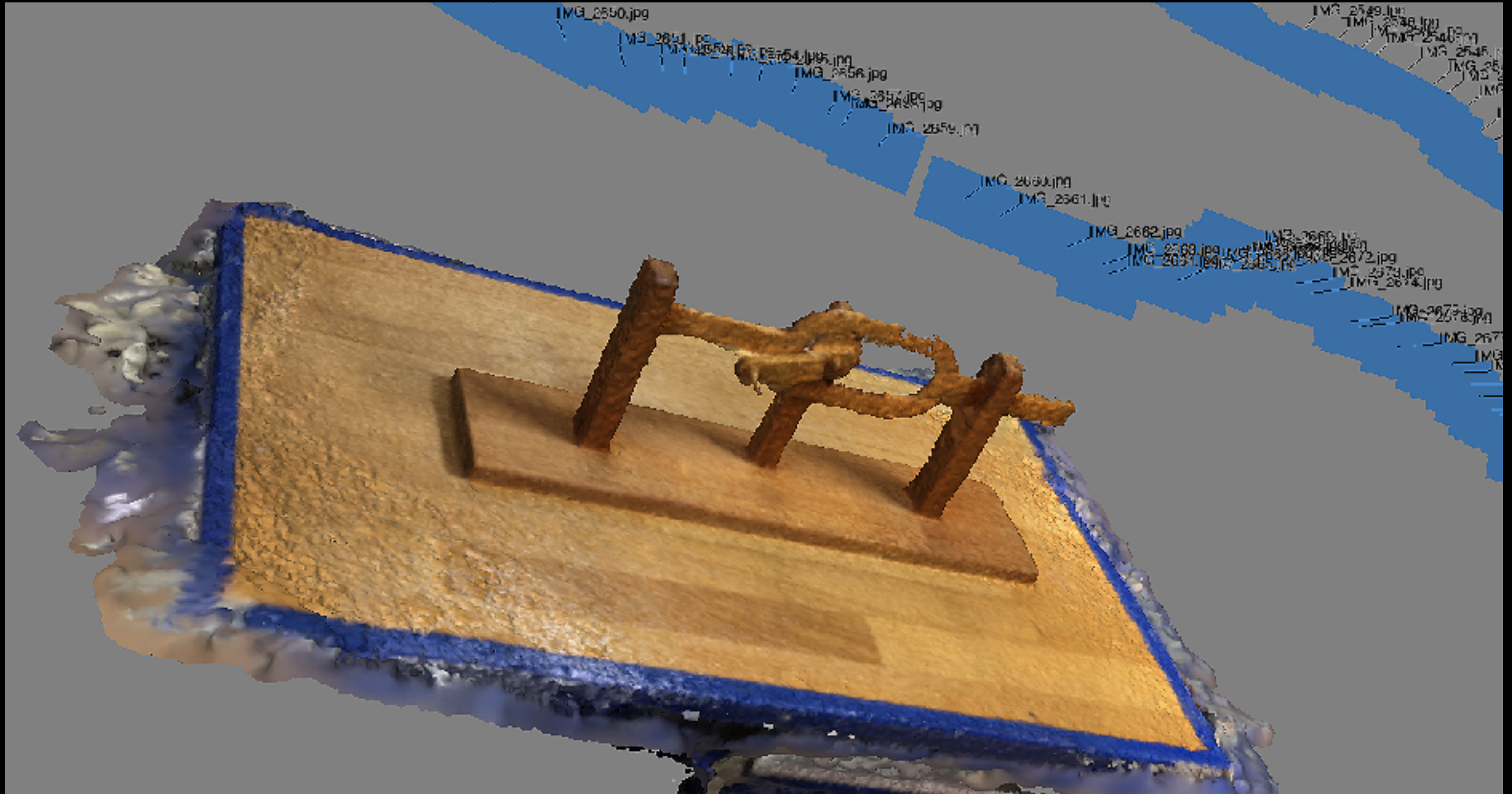
# A. Stupid Scanning

**123D Catch is a free app that lets you  
create 3D scans of virtually any object.**

Available on:







Simple scanning of one wooden model from Polhem's mechanical alphabet—using an iPhone and the software Agisoft Photoscan. The IMG.jpg-markers indicate where photographs were taken.

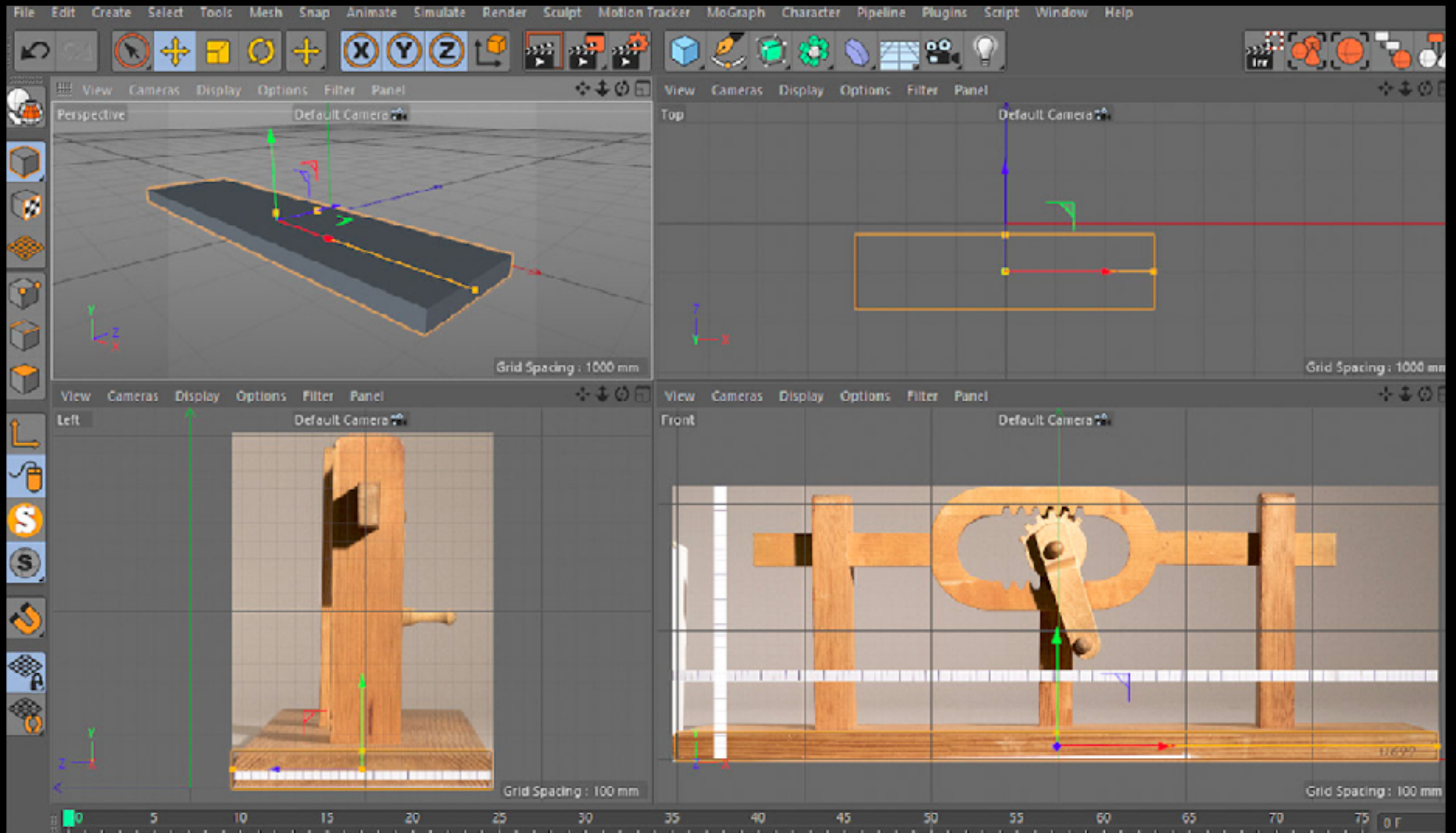
[polhem\\_modell.psz](#)



## II. Computer Animated Models

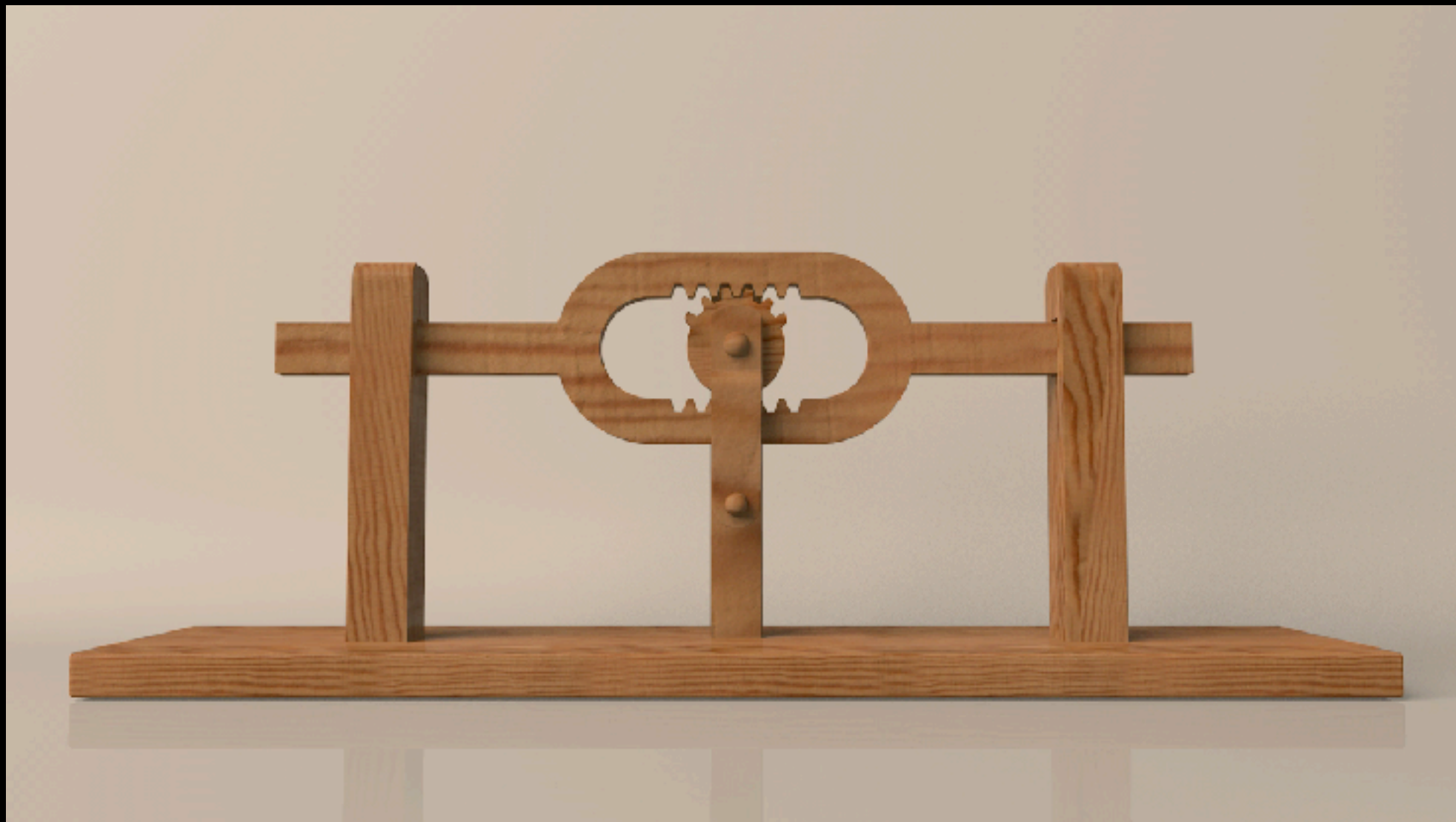
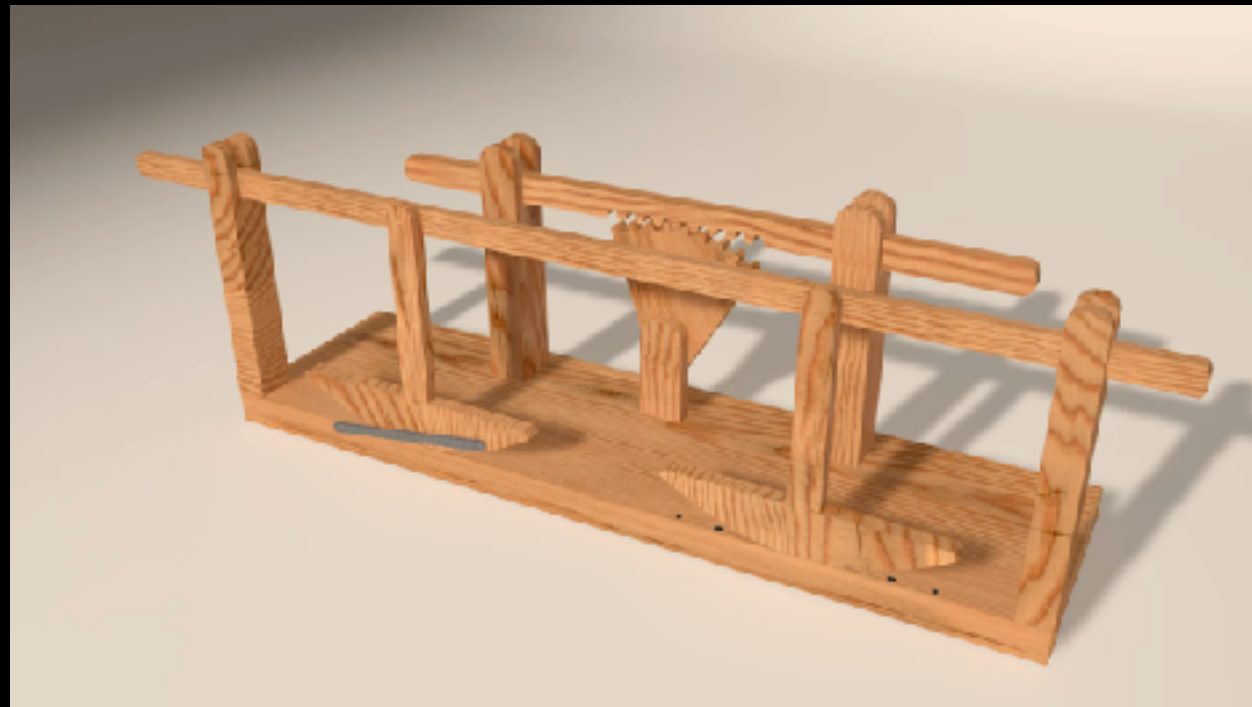
We collaborated with the professional Swedish animator Rolf Lindberg who computer-animated some models (rather than scanning them).





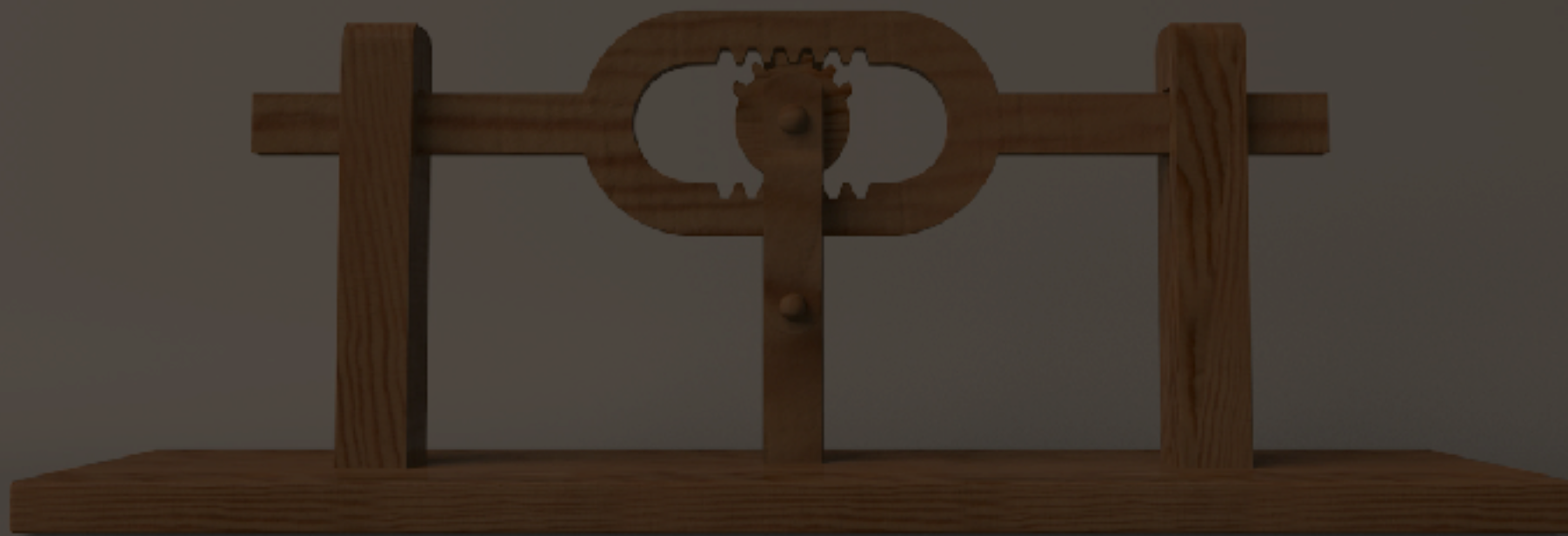
Rolf Lindberg simulated a model from Polhem's mechanical alphabet—by way of a few photographs—and constructed a brand new virtual object in the software Cinema 4D.







The difficulty in rendering Polhem's models based on "technical rigour in digital heritage visualisation"—to quote the London charter on 3D heritage—became especially problematic regarding animations of **model movement**.

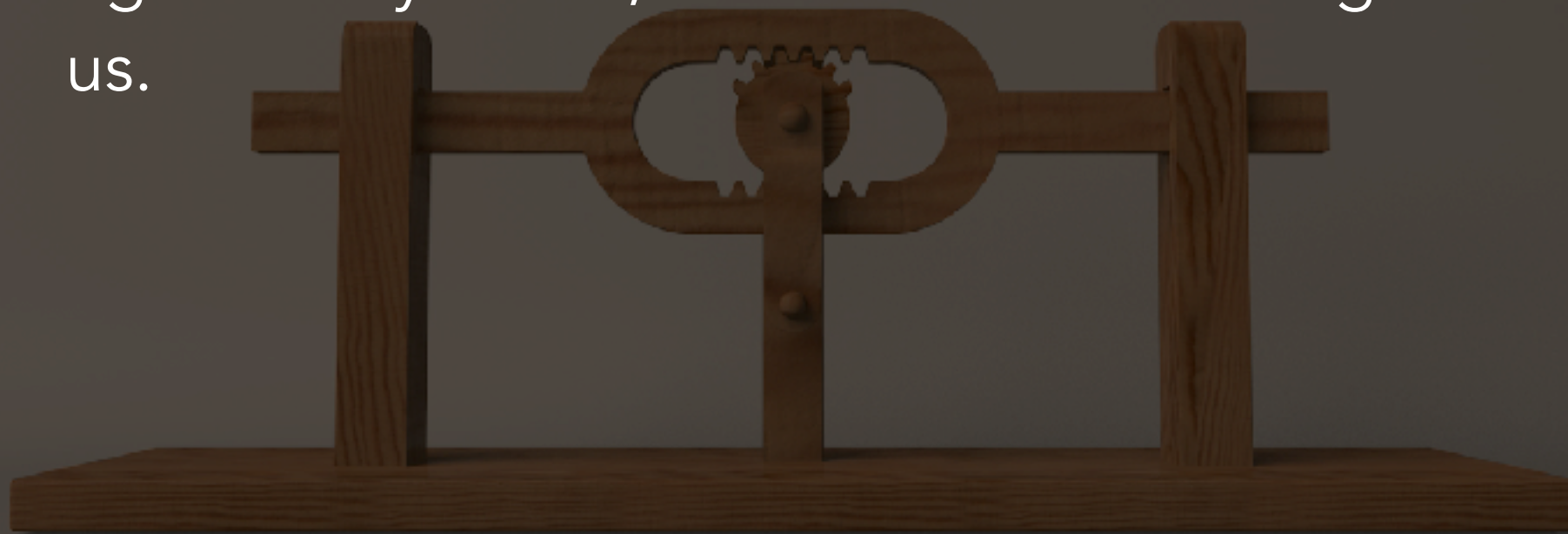




Naturally, computer code could make all models run completely smooth. In one original model, however, a cogwheel caused a lot of friction; the model was built by wood after all.

We hence asked Lindberg **to make friction more noticeable** in his animations when the cogwheel moved (which he did).

In fact, the tricky issue of how to represent friction in a technical rigours way in 3D, became an interesting research question for us.

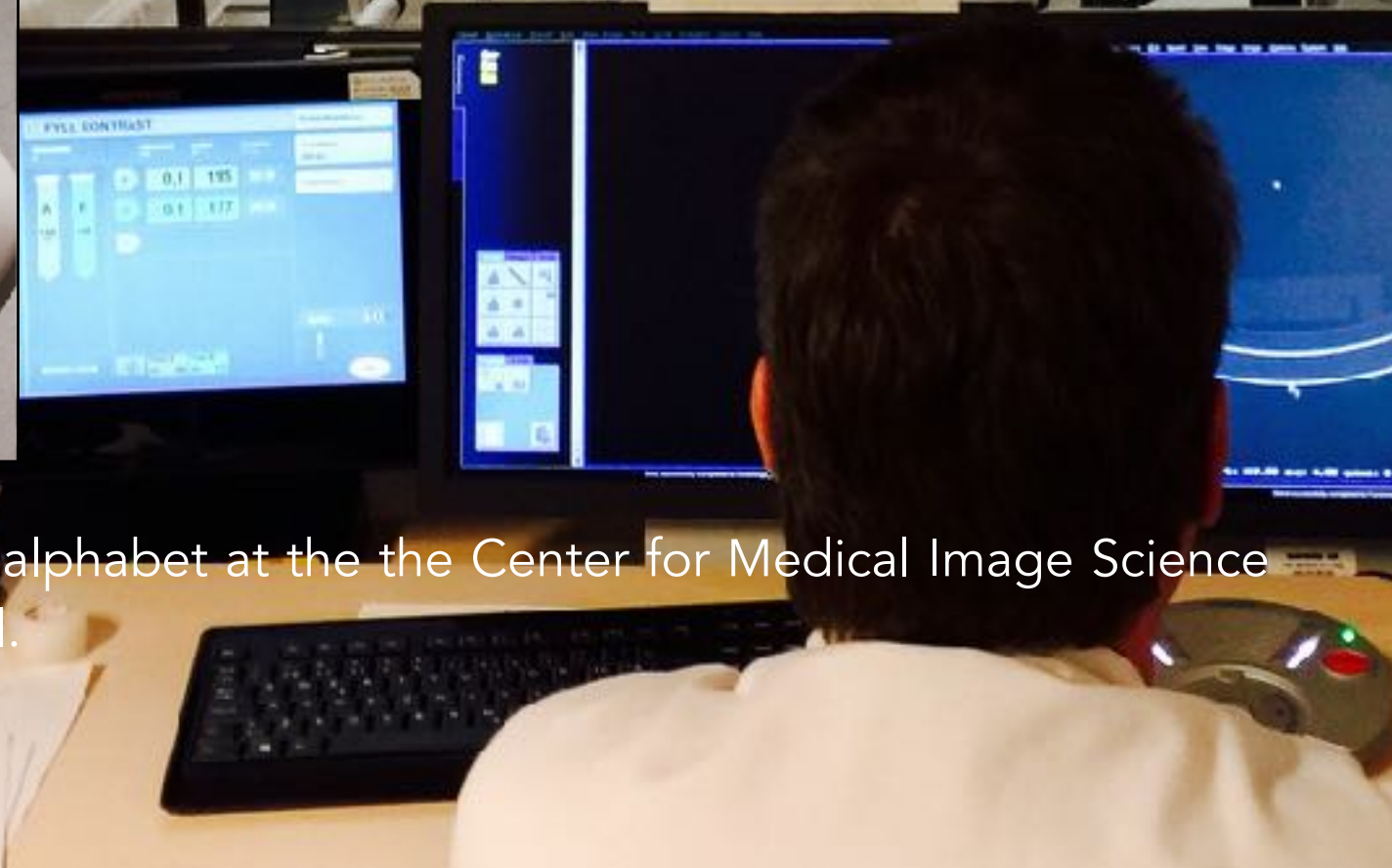
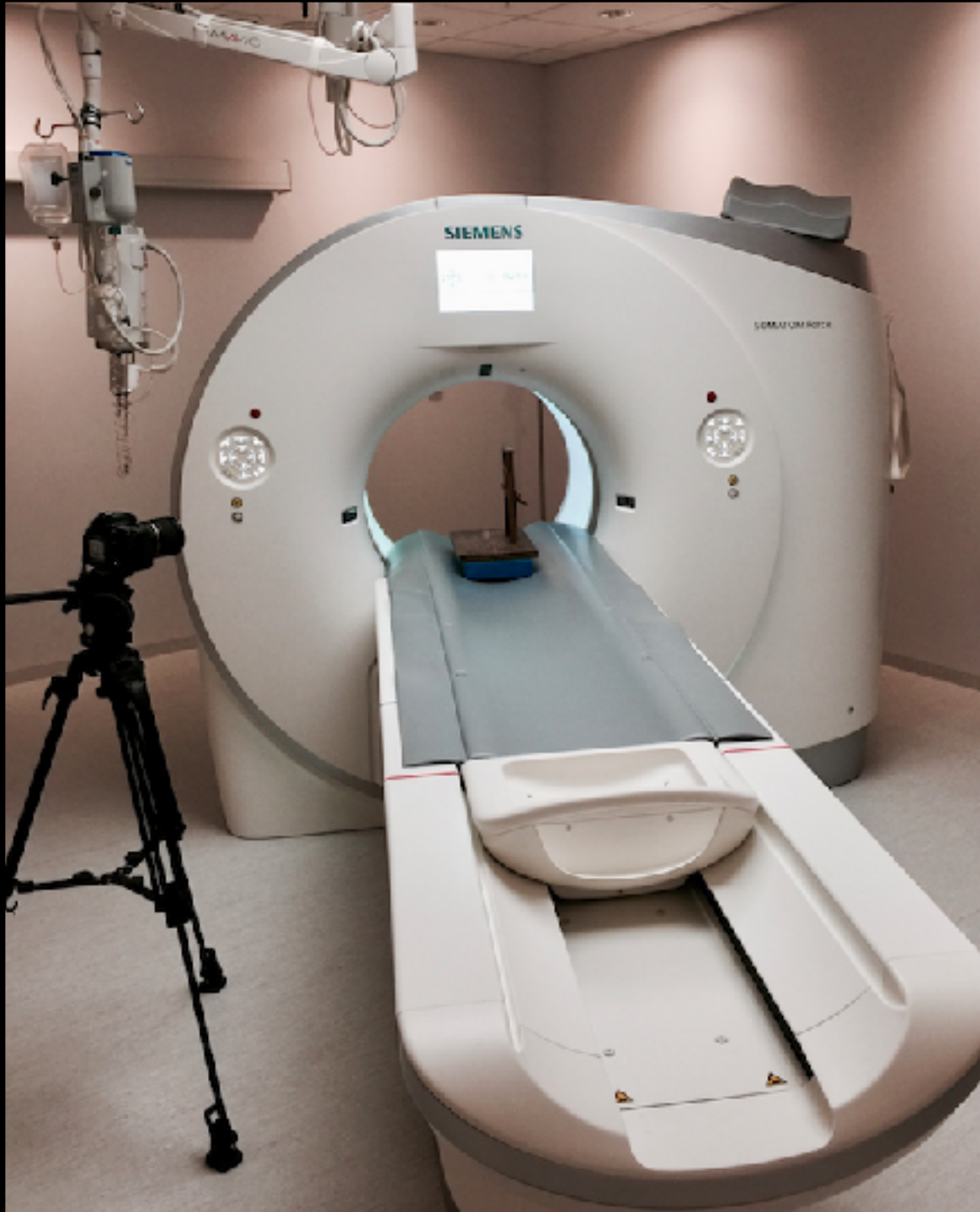




### III. CT-Scanning Models

We CT-scanned some models—that is, X-ray computed tomography—at Linköping University Hospital in a collaboration with the Center for Medical Image Science and Visualization.



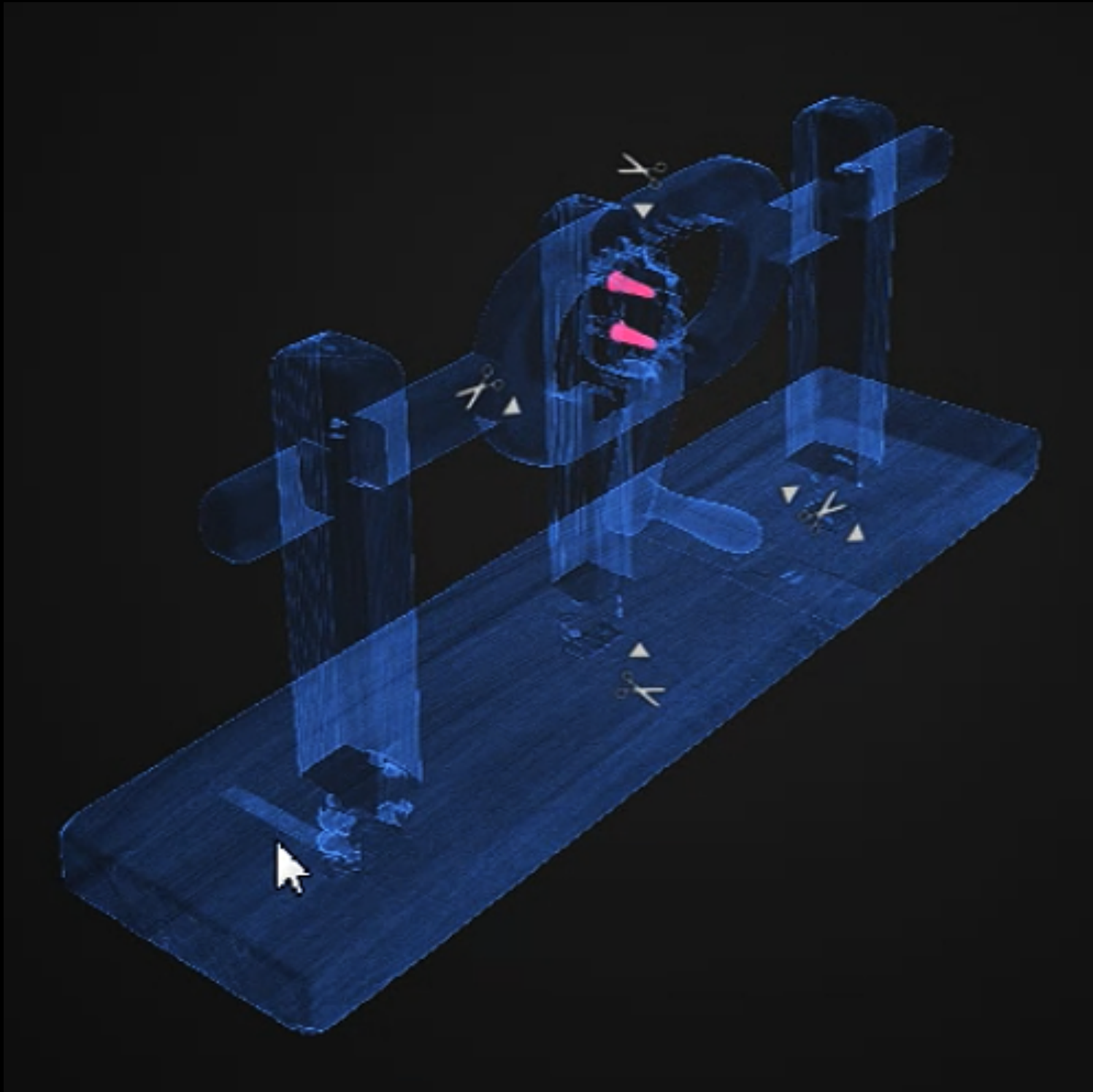


CT-scanning a model from Polhem's mechanical alphabet at the the Center for Medical Image Science and Visualization at Linköping University Hospital.





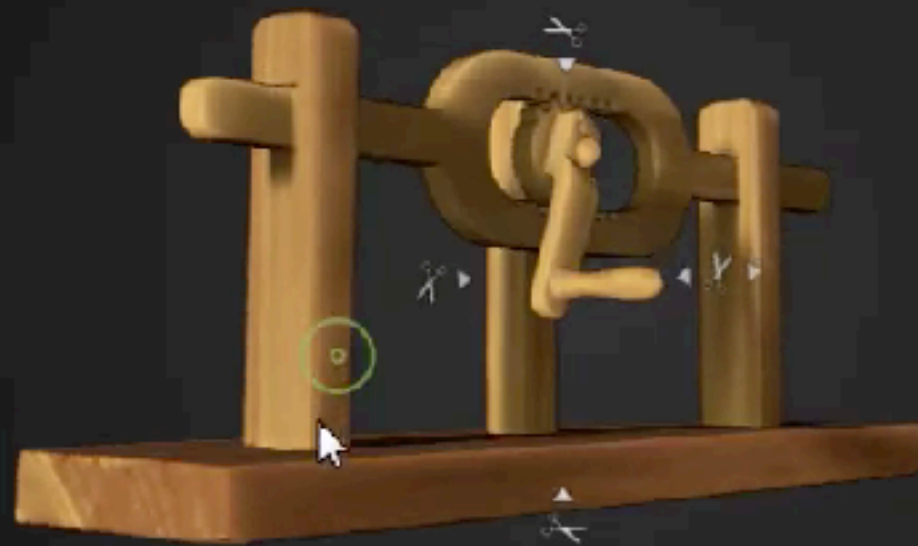




Inside Polhem—CT-scanning a model in collaboration with the company **Interspectral** made it possible **to see inside models** without breaking them.



1359



Solid Wood

Shell & Metal





## IV. Student Scanning

We collaborated with two students that—training to become 3D technicians—who used **photogrammetry to make 3D visualisations of all models.**



# Student photogrammetry





## Visual Models of the Past

The purpose of these try-outs of multiple scanning procedures—or perhaps models of models—was **to raise awareness within the heritage domain** that 3D digitisation and visualisations can be done in various ways.



## Visual Models of the Past

It might not come as a surprise, and the London charter on computer-based visualisation recognises a wide range of available methods. Yet it is quite another matter for a heritage institution to have the ability **to practically test differences in 3D procedures** and results.



## Visual Models of the Past

The **specificity of 3D digitisation depends on factors as selected materials, scanning, rendering and modeling procedures**—not to mention funding. It goes without saying that the contrast between self-scanning Polhem's models and CT-scanning them is foremost one of money.



## Visual Models of the Past

Using Polhem's mechanical alphabet as a case, different 3D digitisation methods will result in **representations that share some attributes with the original models**—but not all of them.

You can, for example, move a 3D model around—but not move its parts. In an animation, on the other hand, all parts move—but you cannot steer movement yourself.

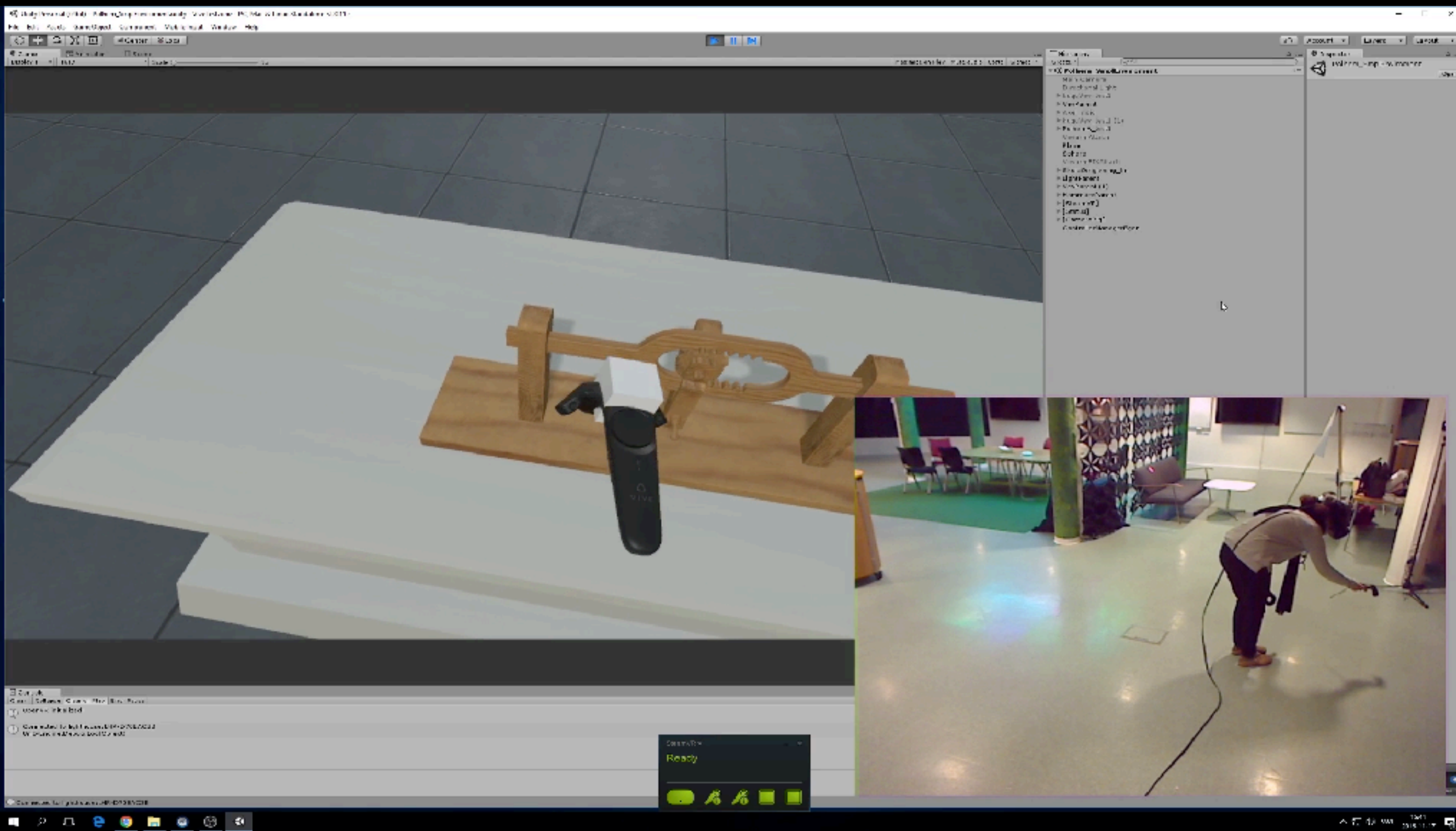


## V. Remodeling the Swedish Model Chamber in VR

At Humlab we hence decided to **re-modelled Polhem's mechanical alphabet by building a virtual reality model**—for HTC Vive glasses with the software Unity—of the Royal Swedish Model Chamber around 1760.



# VR simulation beta









1. Introduction

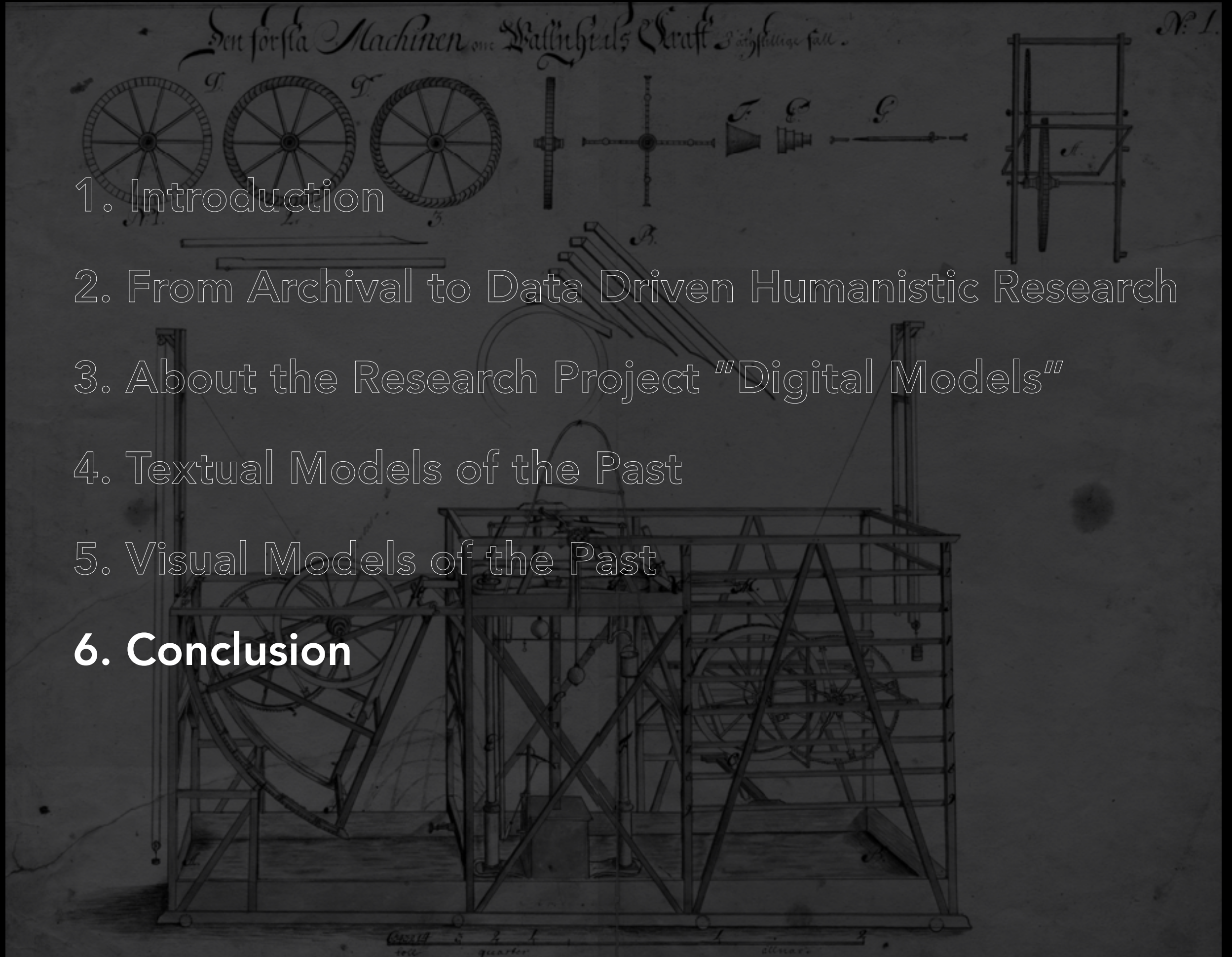
2. From Archival to Data Driven Humanistic Research

3. About the Research Project "Digital Models"

4. Textual Models of the Past

5. Visual Models of the Past

**6. Conclusion**





## Conclusion

As long as the library or “the archive” was considered a place for historical scholarly practices—that is, essentially a repository of a static set of documents against which history refined itself—**digital resources** were foremost a way or an instrument that **provided access** to the past.



## Conclusion

However, all forms of digitisation transforms “the archive”—in quite different ways—and makes it **processable and calculable**.

Documents and resources then become **constitutive parts** of the digital repository. They also become **subject to analysis in their own right**—as James Mussel and others have argued—and they can be analysed anew with the help of **novel tools or applications**.



## Conclusion

I would argue that if the **epistemic foundation of “the archive” is altered**—then **scholarly historical practices needs to change too**. Not all—but some.



## Conclusion

**Modeling the past**—in textual, visual or algorithmic ways—is one way to understand and scholarly work with both the science of history as well as the history of science and technology (in my case) from a somewhat different perspective, where the history of technology hopefully appears in a new light.

3D digitisation can, for example, be done through various **modeling procedures**, even if **textual and algorithmic models** on big data sets needs repeated computational training.



## Conclusion

I would, however, finally also like to **stress the advantage** for the humanities—and in particular the digital humanities—to be able to work with the more general concept of **scientific modeling**.



## Conclusion

I would, however, finally also like to **stress the advantage** for the humanities—and in particular the digital humanities—to be able to work with the more general concept of **scientific modeling**.

It might help us to **learn more**.



– kiitos!

