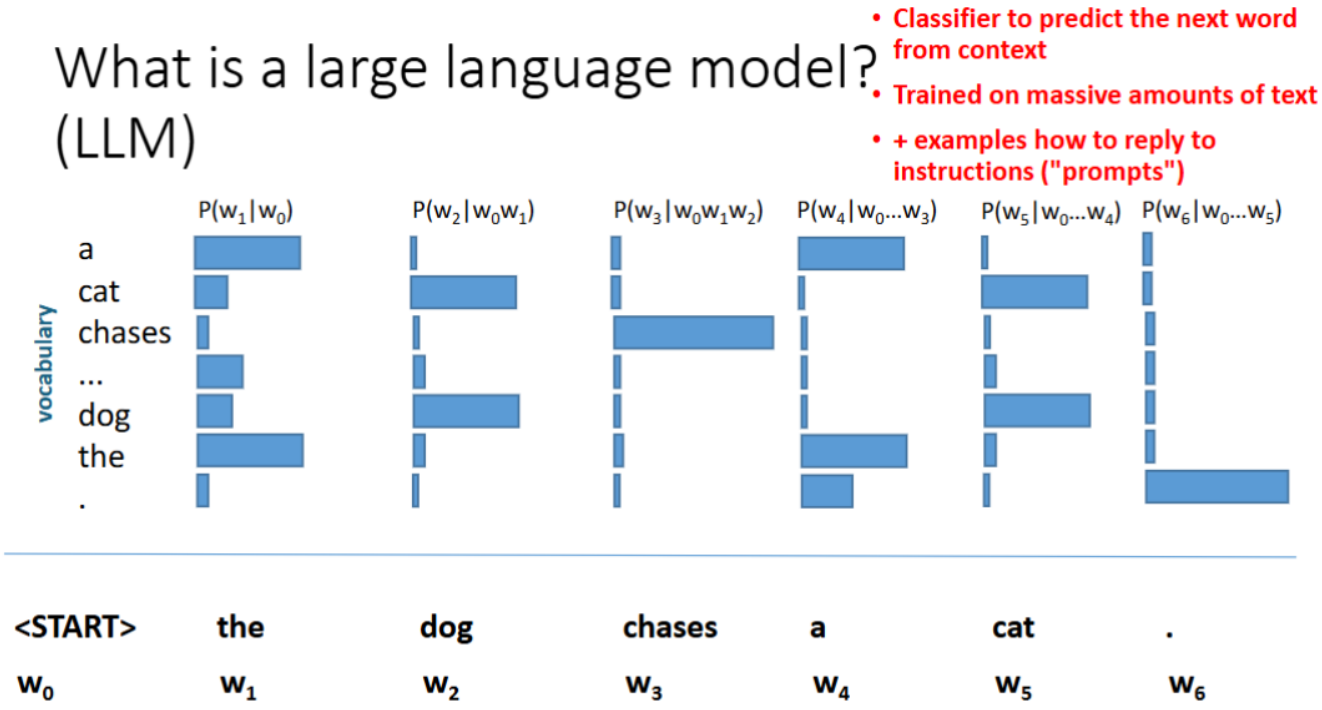
Notes from the presentation "Large language model analysis and applications in Digital Humanities" by Benjamin Roth, Loris Schoenegger, and Vanja Karan from the University of Vienna.

1. Introduction to Large Language Models (LLMs)

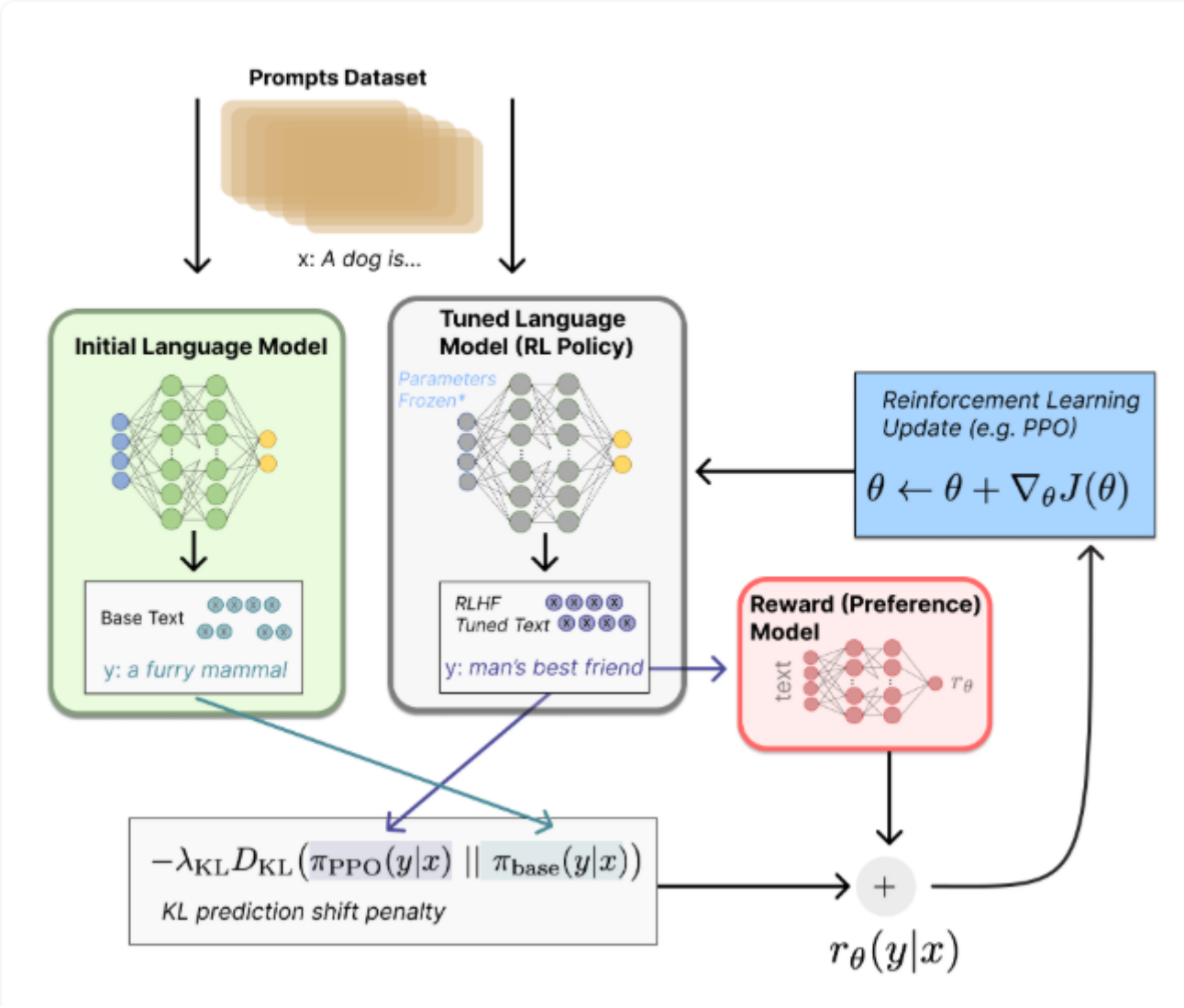
What is an LLM?

- An LLM is an artificial neural network trained on massive amounts of text to predict subsequent text based on a given context (a prompt).
- It works by predicting one word at a time, with each new word being added to the context for the next prediction.
- The process can be initiated with a special `<START>` symbol to generate text from scratch.
- LLMs are essentially classifiers that calculate the probability of the next word given the preceding words (e.g., $P(w_{n+1}|w_{0}...w_{n})$).



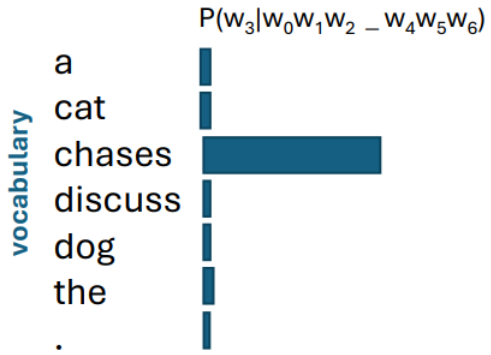
How LLMs Learn

- **Training:** LLMs are trained by initializing their parameters randomly and then repeatedly presenting them with text contexts. The model predicts the next word, and its parameters are adjusted to increase the probability of the correct, observed word. This is done for billions of examples.
- **Pattern Recognition:** Through training, LLMs learn to recognize patterns in language, including groups of similar words and larger linguistic structures.
- **Instruction Tuning:** Modern models like ChatGPT undergo further training to follow instructions. This involves:
    1. Initial next-word prediction training.
    2. Learning from a large set of example instructions and their answers.
    3. Learning from human feedback to ensure answers are helpful, harmless, and honest, a process known as **Reinforcement Learning from Human** Feedback (RLHF).



Types of Language Models

- **Auto-regressive Language Models (ARLM) or Causal Language Models:**
  - These models predict text from left-to-right.
  - They are primarily used for *generating* text.
  - Examples include GPT, GPT-2, and ChatGPT.
  - Better than MLM for **generating text**
- **Masked Language Models (MLM) or Autoencoding models:**
  - These models predict a "masked" or hidden word based on its surrounding context (words before and after it).
  - They are primarily used for *analyzing and categorizing* existing text by learning contextualized vector representations.
  - Examples include BERT, RoBERTa, and ALBERT.



<START>	the	dog	<MASK>	a	cat	.
w_0	w_1	w_2		w_4	w_5	w_6

2. LLM Evaluation

Evaluating LLMs is crucial to understand their capabilities and limitations. Evaluation can be categorized by *what* is being evaluated and *how* it is evaluated.

How to Evaluate

- **Automatic Evaluation:**
  - Involves comparing the LLM's output to a human-written "ground truth".
  - This works well for questions with clear, categorical answers.
  - For tasks with more variable answers, metrics like BLEU or BERTScore are used to measure similarity.
  - A key challenge is that an LLM evaluator might favor answers similar to its own style, regardless of quality.
- **Manual (Human) Evaluation:**
  - Human annotators judge the LLM's output based on various dimensions like fluency, accuracy (veracity), and safety.
  - For creative tasks, humans might rate aspects like grammaticality, cohesiveness, likability, and relevance to the prompt.

What to Evaluate

- **Linguistic & Cognitive Capabilities:** Testing performance on linguistic tasks (syntax, semantics) and tasks that might suggest understanding, like Question Answering or the Winograd Schema Challenge.
- **Psychological Experiments:** Adapting experiments from cognitive psychology to test LLM behavior.
- **Toxicity and Bias:** Checking for unintended behaviors like perpetuating stereotypes, spreading misinformation, or revealing private information.
- **Typical User Queries:** Evaluating performance on common tasks requested by users, such as generation, brainstorming, summarization, and rewriting.

Evaluation Benchmarks and Methods

- **Chatbot Arena:**

    - A platform where users compare outputs from two anonymous LLMs and vote for the better one.
    - Models are ranked using an Elo rating system.
    - **Advantages:** Uses real human judgments on a wide range of tasks.
    - **Disadvantages:** Results can be influenced by user preferences, potential for manipulation exists, and it's not suitable for direct feedback during model development.
- **MMLU (Massive Multitask Language Understanding):**
    - A dataset of over 15,000 multiple-choice questions across 57 subjects, from humanities to science.
    - **Advantages:** Covers a wide range of topics and can track the quality of new models.
    - **Disadvantages:** Multiple-choice format doesn't test text generation, it's single-turn (no conversation), and there's a risk of data leakage (questions being in the training data).
- **Multi-modal Evaluation:**
    - Uses datasets that combine different data types, like text and images, to evaluate models that can process them. The Minecraft Dialogue Corpus is an example where a model must understand a conversation in the context of a visual game state.

While leaderboards and benchmarks offer a way to rank language models based on general capabilities, licensing, and size, they are insufficient for evaluating performance on your specific task. This is especially true in complex systems like retrieval-augmented generation, where the quality of the documents and the retrieval process heavily influence the final outcome.

3. Explainability for LLMs

Explainability seeks to understand *why* a model makes a certain decision.

- **Interpretability vs. Explainability:**
    - **Interpretability** is a property of the model itself; the model is transparent and can be understood by humans (e.g., a simple decision tree). LLMs are generally not interpretable.
    - **Explainability** involves calculating and presenting the most important factors that led to a model's decision.
        - **Faithfulness vs. Rationalization:**
            - An explanation is **faithful** if it accurately reflects the model's underlying causal process.
            - A **rationalization** is an explanation that seems plausible but is not faithful to the model's actual reasoning. Self-explanations from LLMs are often rationalizations and should not be trusted to understand the model's internal logic.

Types of Explanations

- **Feature-based:** Identifies which input features (e.g., words) had the most impact on the output. Methods include LIME and SHAP.
- **Example-based:** Identifies which training examples most influenced the model's output.
- **Mechanistic:** Summarizes the causal dependencies within the model itself.

Evaluating Explanations

Explanations are evaluated on several dimensions:

- **Faithful:** The explanation truly reflects the model's reasoning.
    - **Example:** If a loan application model denies a loan due to "low income," a faithful explanation confirms that income was indeed the primary factor the model used, not a simplified guess while the real reason was something else, like "short credit history."
- **Stable:** The explanation doesn't change wildly with tiny, irrelevant changes to the input.
    - **Example:** An explanation for why an image is classified as a "cat" should highlight the cat's features. If slightly brightening the image causes the explanation to shift from highlighting the "pointy ears" to the "background color," it is not stable.
- **Useful:** The explanation helps you understand the model and take action.
    - **Example:** An explanation like "output neuron #3 fired with an activation of 0.87" is not useful. A useful explanation would be, "The model identified this as a fraudulent transaction because the purchase amount is unusually high for this time of day."

Performance model and task dependent!!

4. LLMs in Digital Humanities

LLMs have a wide range of applications in the field of Digital Humanities:

- **Text Analysis and Interpretation:** Identifying themes, sentiment, stylistic features, and attributing authorship.
- **Data Mining:** Analyzing large text corpora to uncover historical patterns in migration, economic trends, and cultural dynamics.
- **Language Translation and Transcription:** Processing large volumes of text that are impractical to handle manually.
- **Digital Archiving and Curation:** Digitizing content and creating new, interactive ways for users to engage with museum exhibits and archives.
- **Data Visualization:** Assisting in the creation of visualizations from textual data.

Notes from the presentation "BALANCING SPECIFIC NEEDS AND LONG-TERM SUSTAINABILITY IN DIGITAL EDITIONS: AN OPEN CHALLENGE" by Beatrice Nava from the University of Vienna.

1. What are Scholarly Digital Editions (SDEs)?

- **Definition**: SDEs are critical representations of historical documents or texts created and guided by digital methods. They are described as highly diverse and fragile resources.
- **Key Characteristics**:
  - **Complex**: They require collaboration between content specialists, programmers, and data visualization experts.
  - **Multi-layered**: The process involves transcription, encoding, and integrating annotations, facsimiles (images), and other media.
  - **Fragile**: They are at high risk of becoming inaccessible over time.

---

2. The Core Challenge: Sustainability

The central problem is ensuring digital editions remain accessible and functional long-term.

- **Primary Threats**:
  - **Technological Obsolescence**: Software and hardware can become outdated. Custom-made software or publication environments create a dependence on software and hardware longevity.
  - **Funding Issues**: Funding is typically short-term and "output oriented," with no plan for long-term financial planning.
  - **Lack of Standardization**: Projects often develop new workflows and practices instead of using standard encoding, making preservation difficult and costly.

---

3. Two Main Approaches to Sustainability

Approach 1: "Go Standard" (Project-Based)

This approach uses widely accepted, open standards to make a specific project more robust.

- **Key Principles**:
  - **TEI (Text Encoding Initiative)**: A standardized vocabulary (in XML) for encoding texts to facilitate exchange and interoperability. It is often used as the basis for scholarly editions.
  - **FAIR Principles**: Ensuring data is Findable, Accessible, Interoperable, and Reusable.
  - **Standard Visualization Tools**: Using open-source platforms to display the edition.
    - **TEI Publisher**: A framework for publishing TEI documents. TeiPublisher is a user-friendly, open-source framework that enables the publication of TEI-encoded documents without needing extensive programming skills. It provides core functionalities like searching and filtering, and allows for customized visualizations based on the specific TEI Processing Model.
    - **EVT (Edition Visualization Technology)**: A tool for creating user-friendly interfaces for digital editions, often showing text and manuscript images side-by-side. The Edition Visualization Technology (EVT) is a user-friendly, open-source tool designed for the

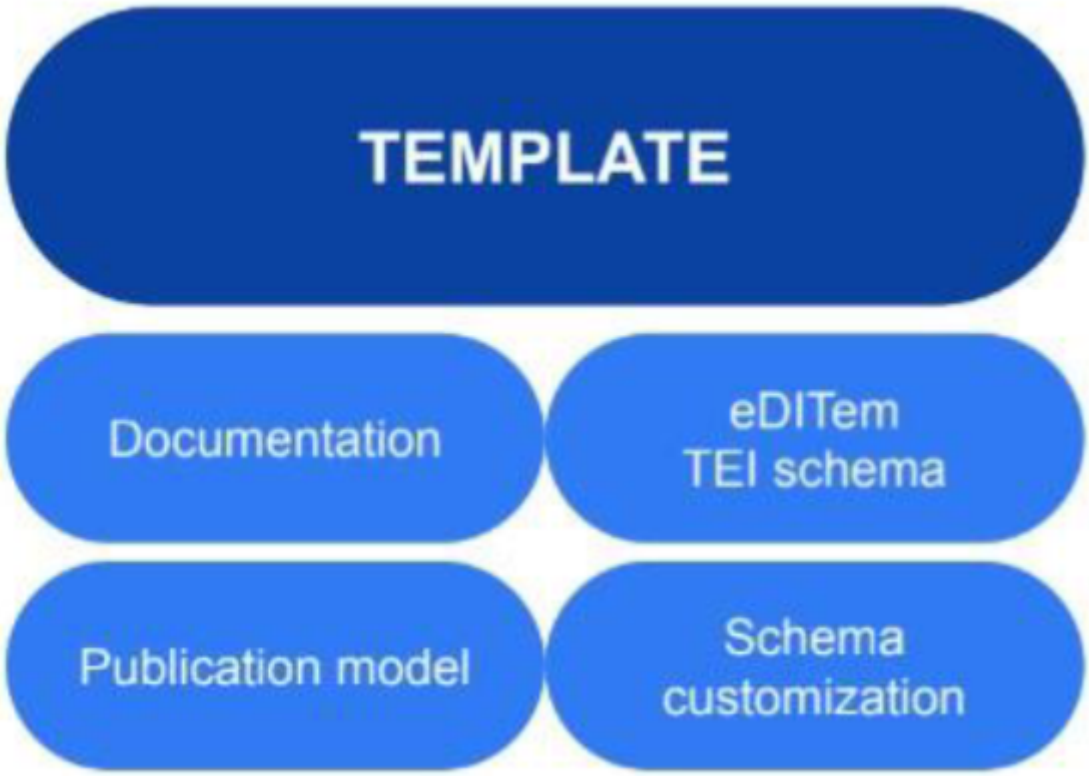
visualization and Browse of TEI-encoded scholarly editions. It supports the side-by-side comparison of manuscript facsimiles and their transcriptions and requires no extensive programming knowledge to use. Additionally, EVT provides built-in tools for searching, filtering, and annotating the textual content of the editions.

- **Case Study: *Leggo Manzoni***
    - **Goal**: Create a digital edition of Manzoni's novel *I promessi sposi*—a foundational text for the modern Italian language—with 40 different critical commentaries.
    - **Challenge**: The commentaries created issues of **overlap** and varying **granularity**, which are difficult to encode in a single file.
    - **Solution**: A **standoff annotation** approach was used.
    - **Sustainability Verdict**: While it uses TEI standards, it relies on a **custom-built infrastructure**, has **no stable team**, and **no long-term funding**.

Approach 2: "Go Bigger" (Infrastructure-Based)

This approach builds shared, centralized platforms to produce and maintain multiple editions efficiently.

- **Analogy**: The Huygens Institute describes digital editions as "newborn puppies" that all need specific, continuous care, which a centralized infrastructure ("zoo") can provide more sustainably.
- **Case Study: eDITem (Huygens Institute)**
    - **Goal**: To create a generic, sustainable environment for TEI-based digital editions through a template approach and a generic publication environment.
    - **Solution: The Template Model**. This model uses reusable templates designed for specific types of documents (like letters or medieval manuscripts) and for common "paratextual" components like bibliographies or introductions.
        - **Composition**: Each template is a package containing four key parts: **Documentation** (instructions on how to encode data), a general **eDITem TEI schema**, a **Schema customization** for the specific document type, and a **Publication model** describing how elements should be published.
        - **Modularity**: An edition is built by combining multiple templates. For example, a complete edition of correspondence could be assembled from a `Letter template` + `Introduction template` + `Biographical template`.
        - **Example (The General Template)**: A base template might include standards for a **facsimile** (images linked page-by-page), the **original text** (both a direct transcription and an edited reading version), **translations**, and **metadata** (date, location, etc.). It also defines allowed annotation types, such as `typednotes` (general remarks), `ogtnotes` (ongoing topic notes), and `notes` (critical commentaries).
    - **Sustainability Verdict**:
        - **Good**: Built on stable servers with a dedicated team and designed for long-term maintenance of many editions "per group".
        - **Bad**: The process is slow, and balancing generic needs with specific project needs requires negotiation and adaptation.



The "Standoff" Approach Explained

The "standoff" approach solves the technical problem of encoding overlapping commentaries by physically separating the primary text from the annotations. Instead of embedding notes directly in the text, they are kept in separate files and linked together.

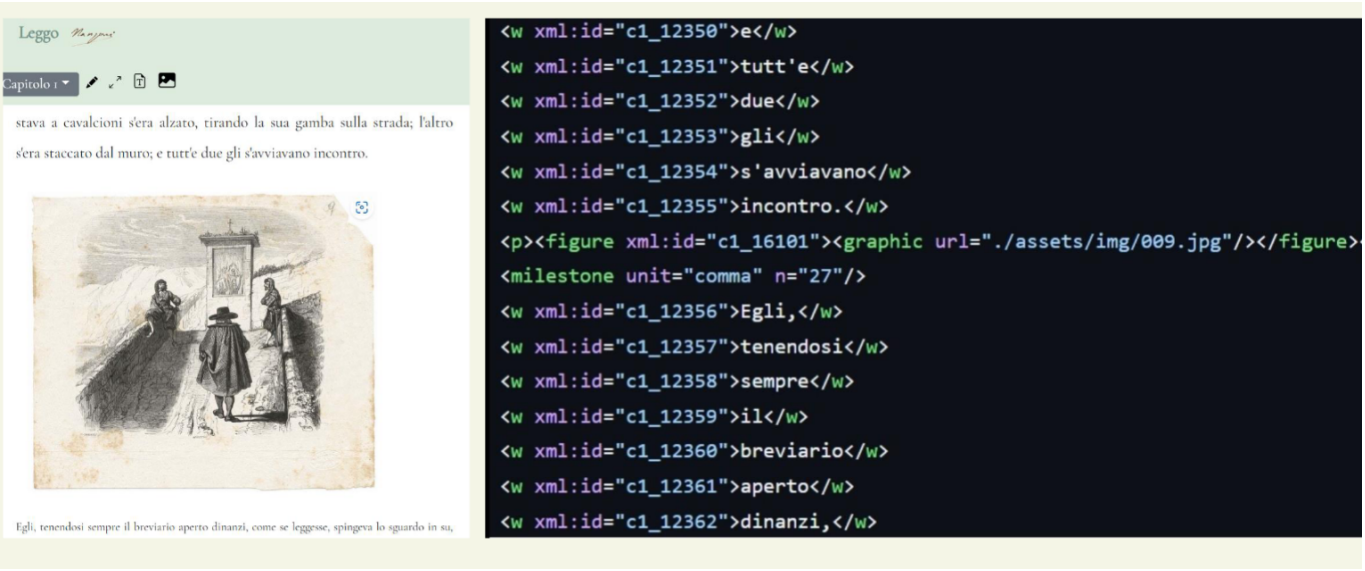
- **Base Text File**: The text of the novel is stored in its own files (e.g., one per chapter). Every single word is automatically tagged as a `<w>` (word) element and given a unique identifier, like `<w xml:id="c1_10002">ramo</w>`. This creates a precise, addressable grid for the entire text.
- **Commentary Files**: Each commentary is stored in a separate file. Within these files, each individual note is encoded in a `<note>` element.
- **Linking the Files**: To link a commentary note to the specific passage it refers to, the `<note>` element uses `@target` and `@targetEnd` attributes.
    - The `@target` attribute points to the unique `xml:id` of the first word in the passage.
    - The `@targetEnd` attribute points to the unique `xml:id` of the last word in the passage.

Here is a simplified example based on the presentation:

```
<note xml:id="BadConf_cap1-n1"
      target="quarantana/cap1.xml#c1_10001"
      targetEnd="quarantana/cap1.xml#c1_10017">

  <ref rend="bold">Quel ramo... monti</ref>: è il ramo verso sud-est...</note>
```

This `<note>` is explicitly linked to the text segment starting at word `c1_10001` and ending at word `c1_10017` in the chapter 1 file. Because the note is in a separate file, it doesn't interfere with any other notes that might refer to the same or an overlapping passage. This makes the system clean, manageable, and scalable.


Leggo
Capitolo I
stava a cavalcioni s'era alzato, tirando la sua gamba sulla strada; l'altro
s'era staccato dal muro; e tutt'e due gli s'avviavano incontro.
Egli, tenendosi sempre il breviario aperto dinanzi, come se leggesse, spingeva lo sguardo in su,
<w xml:id="c1_12350">e</w>
<w xml:id="c1_12351">tutt'e</w>
<w xml:id="c1_12352">due</w>
<w xml:id="c1_12353">gli</w>
<w xml:id="c1_12354">s'avviavano</w>
<w xml:id="c1_12355">incontro.</w>
<p><figure xml:id="c1_16101"><graphic url="./assets/img/009.jpg"/></figure>
<milestone unit="comma" n="27"/>
<w xml:id="c1_12356">Egli,</w>
<w xml:id="c1_12357">tenendosi</w>
<w xml:id="c1_12358">sempre</w>
<w xml:id="c1_12359">il</w>
<w xml:id="c1_12360">breviario</w>
<w xml:id="c1_12361">aperto</w>
<w xml:id="c1_12362">dinanzi,</w>

Notes from the presentation "The Good Life Digital and the Power to Shape It" by rich Prem from TU Wien.

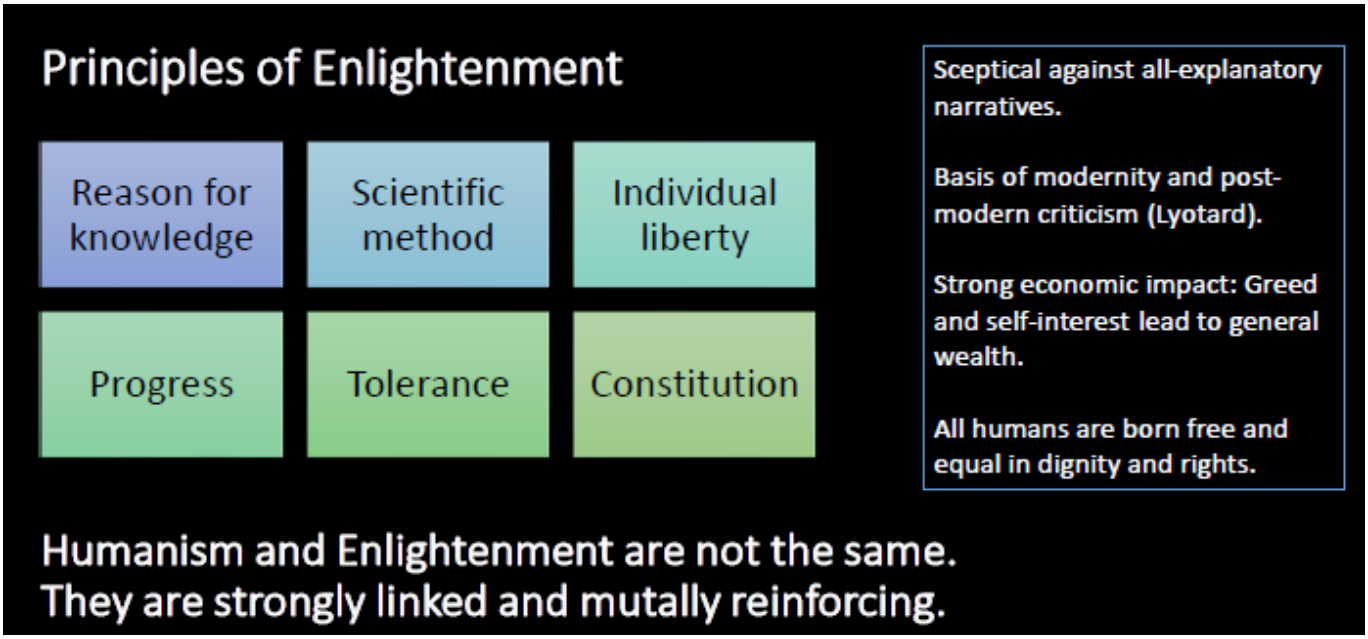
"The Good Life Digital and the Power to Shape It"

Part I: The Good Life Digital

---

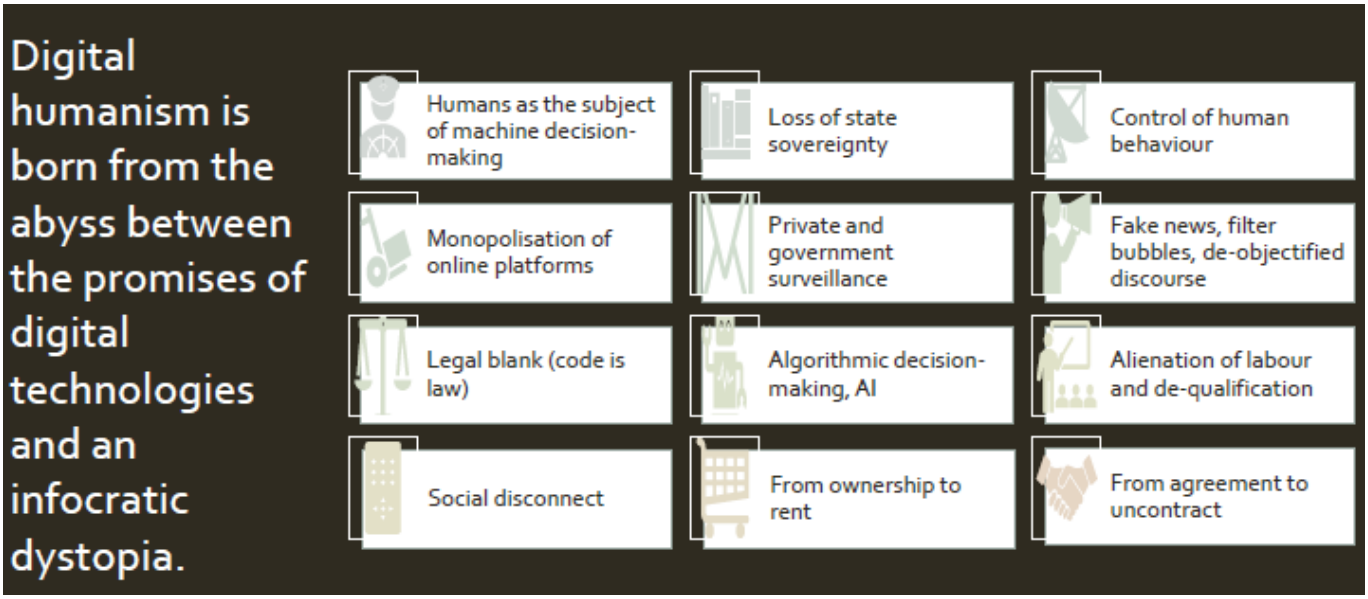
1. Philosophical Foundations

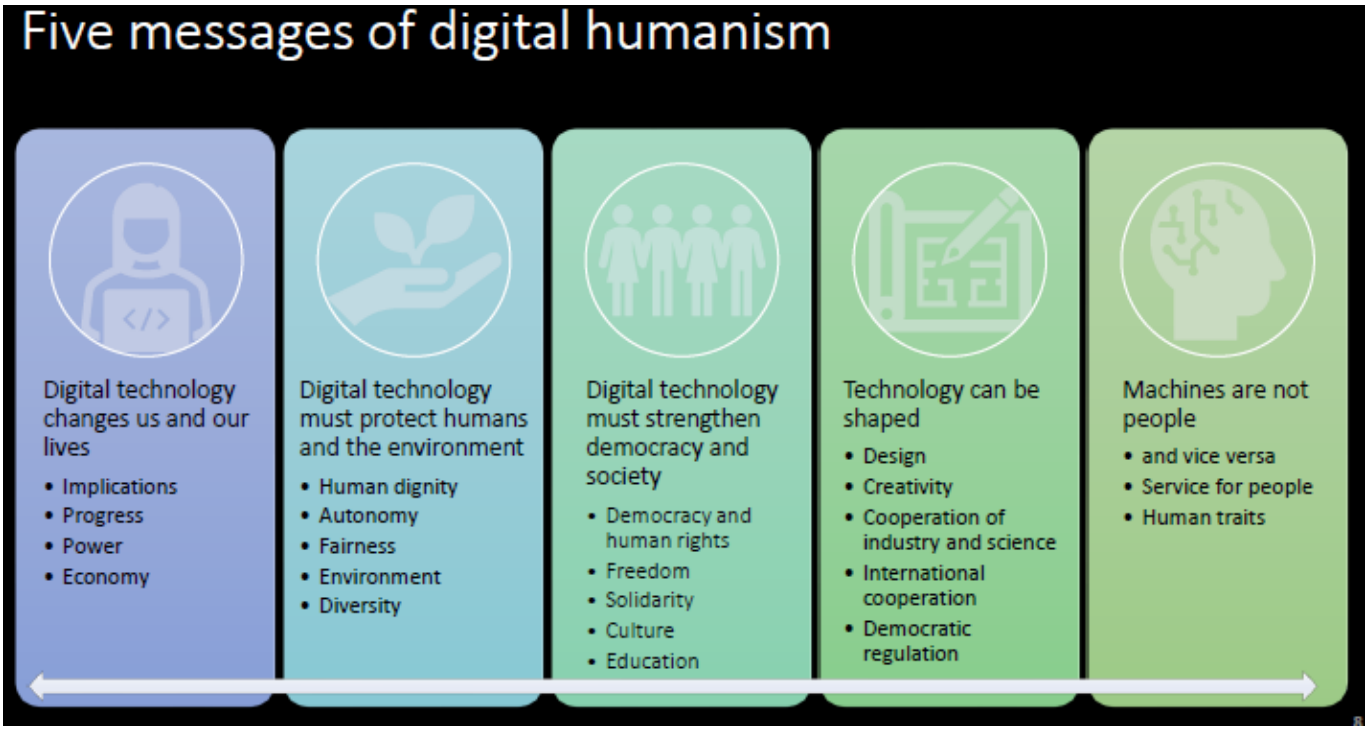
- **The "Good Life"**: The lecture begins by invoking Aristotle's concept of ethics, where the goal of all actions is happiness realized in a "good life". This is defined as an "activity of the soul according to goodness" over the course of a whole life.
- **Humanism & Enlightenment**:
    - The foundation of humanism is Protagoras's idea that "Man is the measure of all things," meaning the perceiving and thinking human is the standard for all things.
    - It is tied to the Enlightenment, which Immanuel Kant defined as "man's emergence from his self-incurred immaturity". This requires the courage to use one's own reason and take responsibility for one's actions.
    - **Key Principles of Enlightenment**:
        - Reason for knowledge
        - Scientific method
        - Individual liberty
        - Progress
        - Tolerance
        - Constitution
- **Critique of Humanism**: The lecture also acknowledges critiques of traditional humanism, such as its human self-centeredness, idealization of antiquity, and colonialist tendencies.

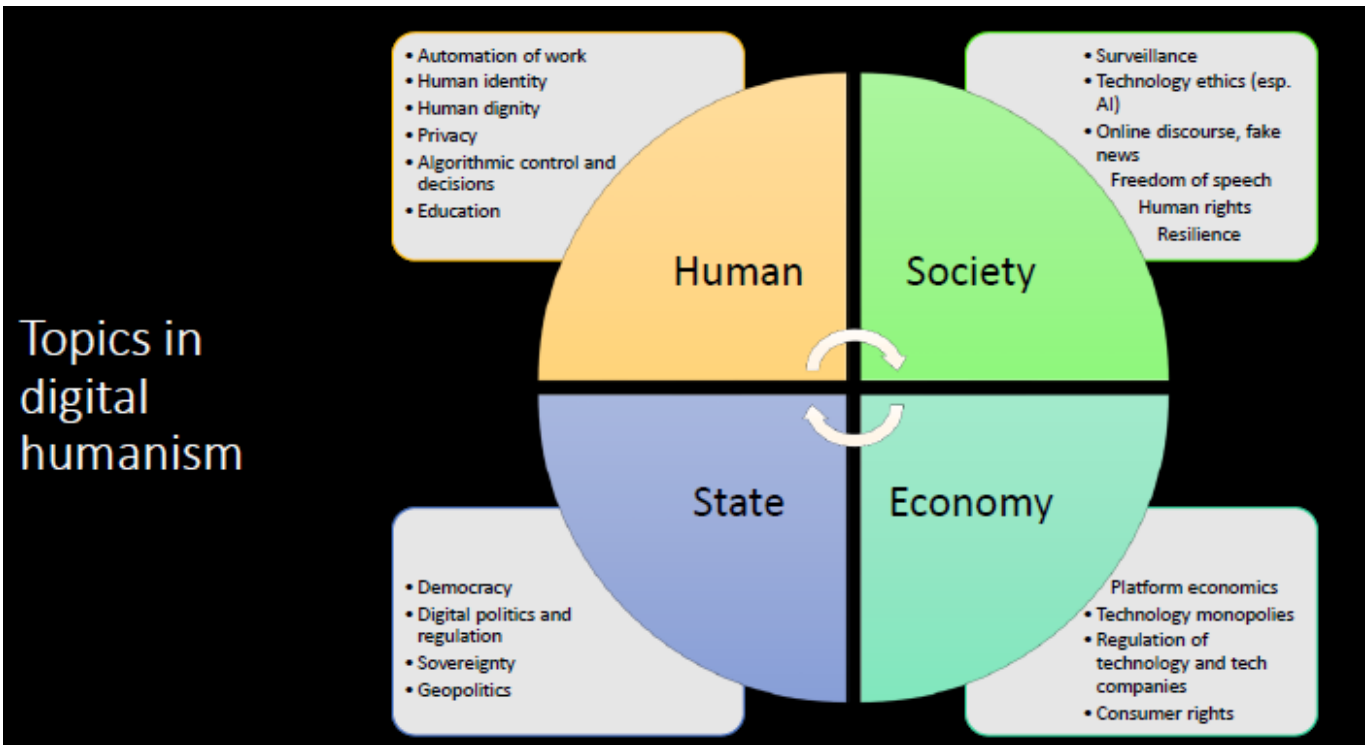


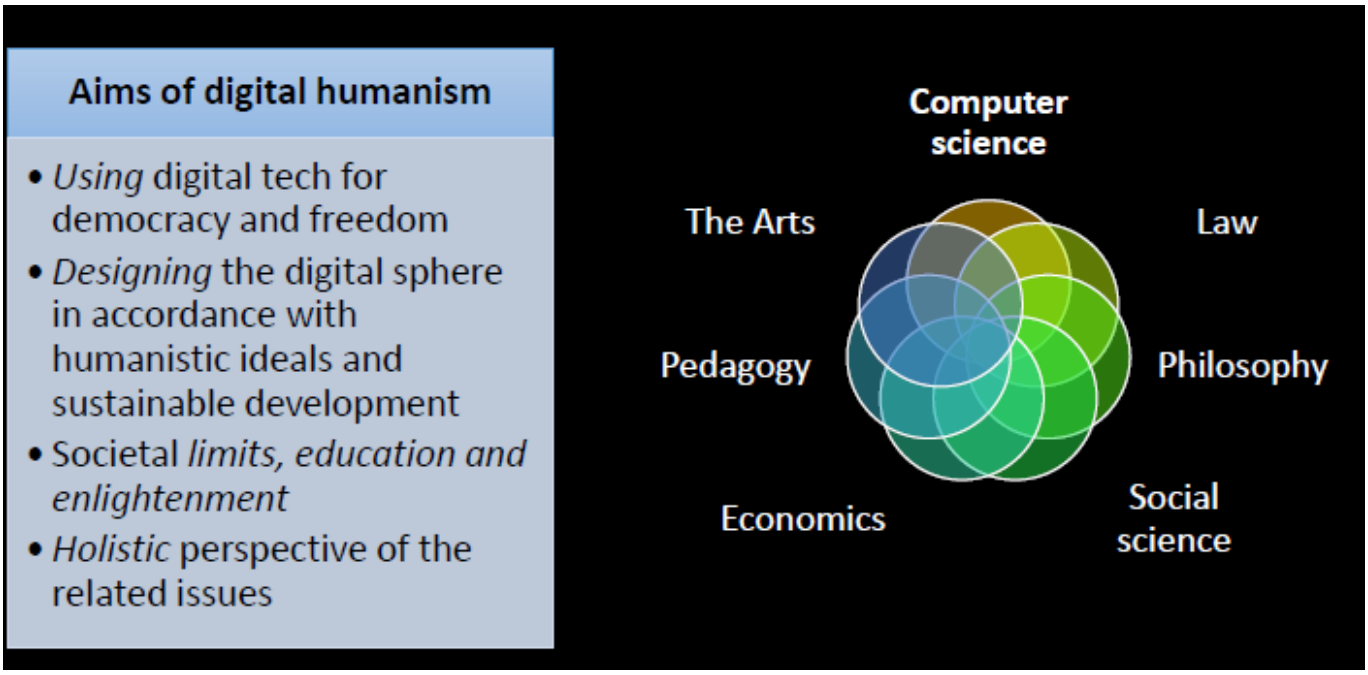
2. The Rise of Digital Humanism

- **The Problem**: Digital Humanism emerges from the gap between the promises of technology and an "infocratic dystopia".
- **Issues in the Digital Age**:
  - Humans becoming subjects of machine decision-making.
  - Monopolization of online platforms.
  - Private and government surveillance.
  - Control of human behavior, fake news, and filter bubbles.
  - Alienation of labor and de-qualification.
  - A shift from ownership to renting.
- **Definition**: Digital Humanism is the effort to strive for human dignity and coexistence in the digital age, shaping technology to support human rights, the common good, and sustainability. It is an engineering and design endeavor to create feasible and worthy visions of a digital future.
- **Aims of Digital Humanism**:
  - Using digital technology for democracy and freedom.
  - Designing the digital sphere in accordance with humanistic ideals and sustainable development.
  - Establishing societal limits, education, and enlightenment.
  - Taking a holistic perspective on all related issues.




Five messages of digital humanism
Digital technology changes us and our lives
• Implications
• Progress
• Power
• Economy
Digital technology must protect humans and the environment
• Human dignity
• Autonomy
• Fairness
• Environment
• Diversity
Digital technology must strengthen democracy and society
• Democracy and human rights
• Freedom
• Solidarity
• Culture
• Education
Technology can be shaped
• Design
• Creativity
• Cooperation of industry and science
• International cooperation
• Democratic regulation
Machines are not people
• and vice versa
• Service for people
• Human traits


Topics in digital humanism
• Automation of work
• Human identity
• Human dignity
• Privacy
• Algorithmic control and decisions
• Education
• Surveillance
• Technology ethics (esp. AI)
• Online discourse, fake news
Freedom of speech
Human rights
Resilience
Human
Society
State
Economy
• Democracy
• Digital politics and regulation
• Sovereignty
• Geopolitics
Platform economics
• Technology monopolies
• Regulation of technology and tech companies
• Consumer rights



3. The Vienna Manifesto & Core Principles

The Vienna Manifesto on Digital Humanism outlines key principles for shaping our digital future:

- **Democracy & Inclusion**: Technologies should be designed to promote democracy and inclusion.
- **Rights & Freedoms**: Privacy and freedom of speech are essential values. Platforms like social media must be altered to better protect these rights.
- **Regulation**:
  - Effective laws must be established based on public discourse to ensure fairness, accountability, and transparency in algorithms.
  - Regulators must intervene to restore market competitiveness against tech monopolies.
- **Human-in-Control**:
  - Decisions with the potential to affect human rights must be made by accountable humans, not left to markets or machines.
  - Automated systems should only support, not replace, human decision-making.
- **Education & Collaboration**:
  - Interdisciplinary approaches are required, breaking down silos between computer science, social sciences, and humanities.
  - Universities have a special responsibility to produce new knowledge and cultivate critical thought.
  - Education on computer science and its societal impact must start early.

Part II: The Power to Shape It

---

1. How Technology Shapes Us

- **Design is Not Neutral**:
  - Simple design decisions in interfaces have a lasting impact on what is considered important. The interface can empower or disempower users.

    - Users are often given the illusion of choice, which is limited by formal constraints (e.g., limited gender options) or "dark patterns" designed to trick them.
- **Disowning Through Software**:
    - There is an "illusion of ownership" when physical products are controlled by software.
    - **The John Deere Case**: The company tried to prevent farmers from modifying the software on tractors they owned (costing over $100,000). In response, farmers began hacking their own equipment using "pirate software" to access diagnostic programs. The case has led to class-action lawsuits.
- **Unilateral Changes**:
    - Companies change terms of use, requiring users to agree to new data collection to continue using a product they already own. Examples include TV manufacturers transmitting viewing habits or vacuum cleaners submitting apartment layouts.
    - This raises questions about the meaning of "agreement" when consent is coerced.
- **Geopolitics**: Digital systems are critical infrastructure for national economies and are now tools of power and targets in conflicts.

2. How to Reclaim Power

- **Set Limits Democratically**:
    - We must lead a democratic discourse to establish rules and limits for technologies that significantly impact our lives.
    - Society, not companies or the market, should make the law.
    - The state must have the authority to enforce these rules everywhere.
- **Recognize Algorithmic Limits**:
    - Algorithms struggle with context and intent. For example, it is difficult for a machine to distinguish between art, medicine, and pornography based on formal criteria like nudity alone.
    - AI can generate harmful or illegal texts, such as instructions for making drugs or encouraging suicide.
- **Be Careful What You Wish For**:
    - A key philosophical question is raised: Should we aim for a society that makes unruly behavior *impossible* (e.g., through perfectly ethical AI), or should we preserve the human "right to violate the rules?".
- **Core Principle: "Machines are not people."**
    - Digital technology should be beneficial for people, not the other way around.
    - The differences between humans and machines should not be blurred.

3. Related Concepts

- **Transhumanism**: The idea of improving the human body with technologies like bio-implants or brain-computer interfaces to overcome deficits like illness and death.
- **Posthumanism**: The idea that an artificial intelligence will eventually overcome and surpass humans, making the human body useless.

Conclusion

The central message of the lecture is that **technology is not a destiny**. It can and must be shaped through democratic processes to align with humanistic values and create a "good life digital."

Notes from the presentation "A digital edition of modern research on ancient texts: Wilhelm Siegling's nachlass" by Bernhard Koller from University of Vienna.

A Digital Edition of Wilhelm Siegling's Nachlass

---

1. Background: The Tocharian Languages & Wilhelm Siegling

- **The Languages**:
  - Tocharian manuscripts, discovered in the Tarim Basin (modern Xinjiang, China) around the turn of the 20th century, represent a previously unknown branch of the Indo-European language family.
  - There are two related languages: **Tocharian A** (East Tocharian) and **Tocharian B** (West Tocharian).
- **The Text Corpus**:
  - The texts are mostly Buddhist literature, with a small number of secular documents like monastic records and caravan passes.
  - The manuscripts are severely damaged and fragmentary; perfectly preserved leaves are very rare.
- **Wilhelm Siegling (1880-1946)**:
  - A co-founder of Tocharian studies, alongside his collaborator Emil Sieg.
  - He published the first grammatical overview (1908), a major edition of Tocharian A texts (1921), a grammar of Tocharian A (1931), and posthumous editions of Tocharian B texts.

Example sentence from Tocharian A

kāsu ñom-klyu tsraṣi-śśi śäk kälyme-ntw-aṃ sätkatär
good name-fame energetic-of ten direction-PL-in spread

The good fame of the energetic ones spreads in ten directions.
(A 1 a1)

Some words with Indo-European etymologies:

- *ñom* 'name' … Latin *nomen*, Greek *onoma*, English *name*
- *klyu* 'fame' … Greek *kleos*, related to German *laut*
- *śäk* 'ten' … Latin *decem*, Greek *deka*, English *ten*
- *kälyme* 'direction' … related to Greek *klima* 'inclination, slope'

---

2. Wilhelm Siegling's Nachlass

Wilhelm Siegling (1880–1946)

- **Born**: 1880 in Erfurt
- **Died**: 1946 in Berlin
- **Field**: Co-founder of **Tocharian studies** (with Emil Sieg)

Education and Early Career

- **1901–1906**: Studied **Avestan**, **Sanskrit**, and **Tibetan** in Berlin
- **1906**: Began collaborating with **Emil Sieg** to decipher Tocharian manuscripts at the *Museum für Völkerkunde*, Berlin

Major Contributions

- **1908**: Published the **first grammatical overview** of Tocharian
    - Introduced the distinction between **Tocharian A** and **Tocharian B**
    - *(Sieg and Siegling 1908)*
- **1915–1918**: Served in **World War I**
- **1921**: Published an edition of **466 Tocharian A fragments** from the Berlin Turfan collection
    - *(Sieg and Siegling 1921)*
- **1931**: Published a **grammar of Tocharian A**
    - *(Sieg, Siegling, and Schulze 1931)*
- **1949, 1953**: Posthumous publications of **633 Tocharian B fragments** from the Berlin Turfan collection
    - *(Sieg and Siegling 1949, 1953)*

- **What it is**: Siegling's academic estate (*Nachlass*), stored at the Georg-August-Universität Göttingen, consists of his research materials. This project was provided with high-resolution scans of the documents.
- **Contents**: The collection includes:
    - Letters (many from Emil Sieg)
    - Postcards (most from Emil Sieg)
    - Letter drafts
    - Siegling's personal notes
- **Why it's Interesting**:
    - **Window into Research**: Sieg and Siegling lived in different cities for most of their 39-year collaboration, so their letters and postcards document the "making of" some of the most important foundational texts in their field.
    - **Attribution of Ideas**: The correspondence reveals how hypotheses were developed and attributed between the two scholars.
    - **Historical Context**: It provides a unique glimpse into the lives and relationship of two academics through two world wars, including the challenges of academic publishing during WWII.
    - **Personal Relationship**: The letters show a close personal relationship, with informal exchanges about money and "feasts".

- **Annotations**: Siegling often made personal annotations and corrections directly on the letters and postcards he received from Sieg, offering further insight into his thought process.

---

3. The Digital Edition Project: Goals & Workflow

Aim: To create a **digital, queryable edition** of Wilhelm Siegling's *Nachlass* and integrate it into a broader research ecosystem.

Main Goals

- Transcribe **all written documents** within Siegling's *Nachlass*
- Create a **TEI-encoded** version of the transcriptions enriched with:
  - References to **scholars**, **literature**, and **linguistic forms**
  - **English summaries** of each document
- **Publish** the data on **CEToM** (Central Asian Texts on the Move) in a **searchable/queryable** format

Status

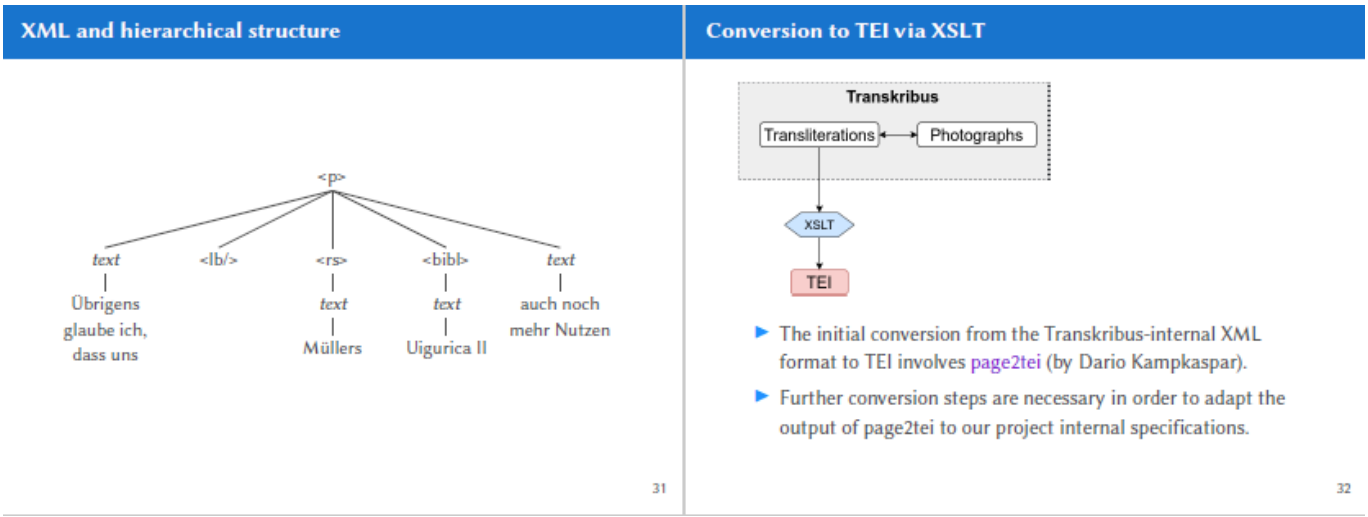
- All **photographs** of the documents are publicly available
- Only a **subset has been transcribed** so far

Step 1: Transcription in Transkribus

- Photographs of the documents are uploaded to **Transkribus**.
- The software performs automatic layout analysis and text recognition (OCR/HTR).
- This automatic output is then **manually corrected**, and entities of interest (like people, places, literary references, and Siegling's own annotations) are marked up with tags.

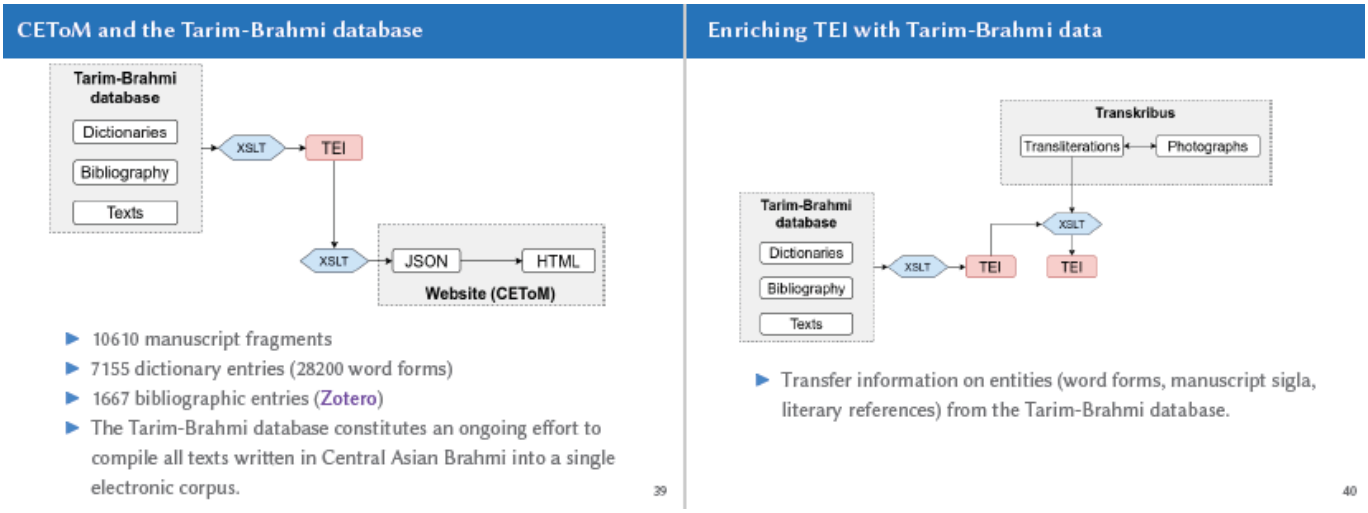
Step 2: Conversion to TEI

- The transcribed and tagged data from Transkribus is converted into a structured **TEI (Text Encoding Initiative) XML** format.
- This is a multi-step process using **XSLT stylesheets** managed by a Python script.
  - An initial conversion is done with a tool called `page2tei`.
  - A custom "TEI Chain" of further XSLT transformations cleans the data, restructures it, expands abbreviations, and adds necessary metadata to create valid, project-specific TEI files.
- **Challenge**: Some letters were folded and used like booklets, meaning the reading order jumps between pages. This is solved either by manually rearranging text after conversion or by using custom tags in Transkribus to automate the reordering during conversion.



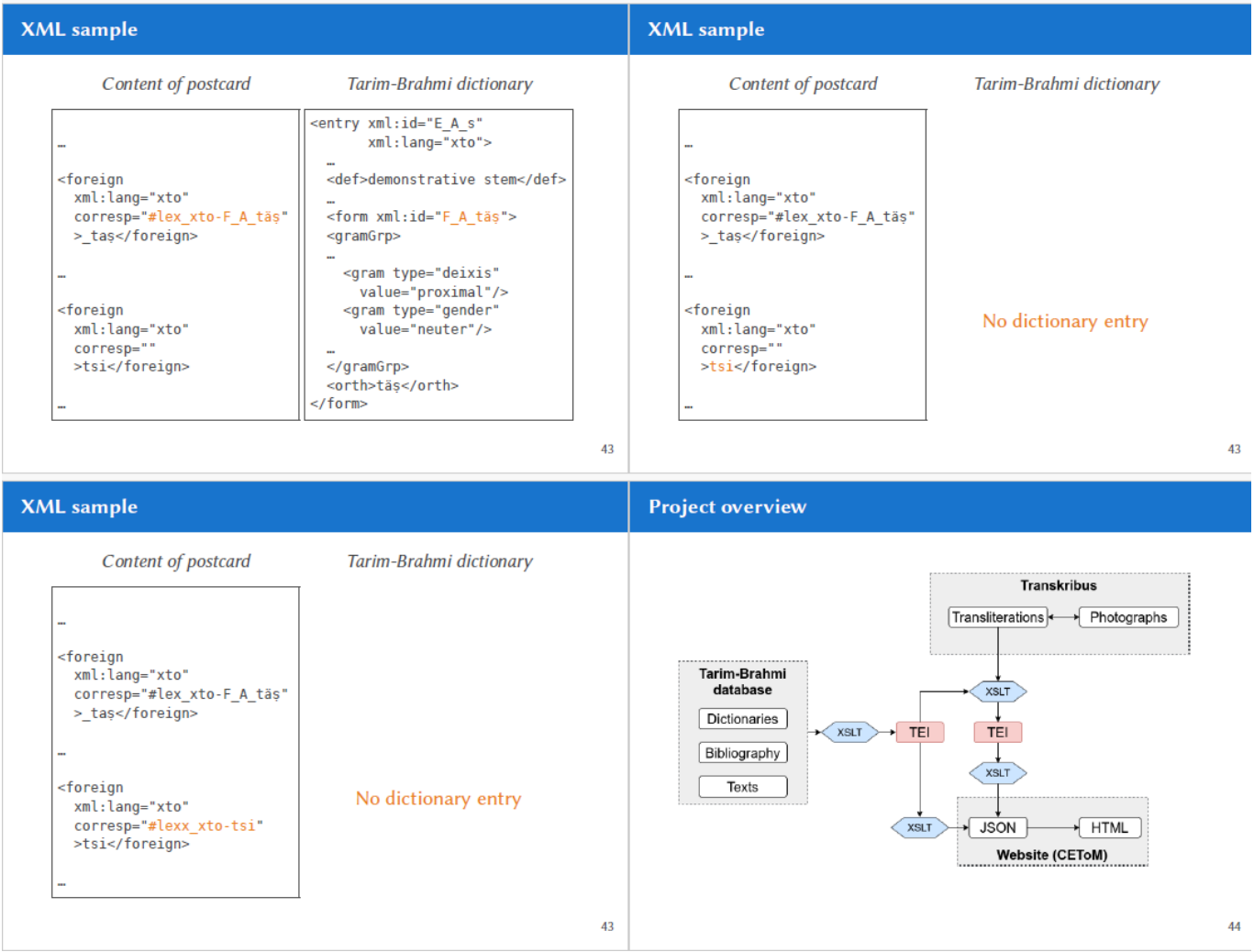
Step 3: Integration with the CETOM Ecosystem

- The project integrates the *Nachlass* data into a pre-existing digital environment for Tocharian studies.
- **CETOM (A Comprehensive Edition of Tocharian Manuscripts)**: An ongoing effort to compile all texts written in Central Asian Brahmi script into a single electronic corpus. It started as an FWF-START project to make edited Tocharian manuscripts publicly available.
- **Tarim-Brahmi Database**: A follow-up project that took CETOM as its starting point. It now includes:
  - 10,610 manuscript fragments
  - A dictionary with 7,155 entries
  - A bibliography with 1,667 entries
- **Enrichment**: The TEI files of the *Nachlass* are enriched by linking them to the Tarim-Brahmi database. For example, when a Tocharian word is mentioned in a letter, a reference is added in the XML that links it directly to its entry in the digital dictionary.



---

4. Publication and Final Output



- **Website**: The final digital edition is published on the CETOM website.
- **Features**:
  - Users can view the original photographs of the documents alongside the transcriptions.
  - The text is searchable.
  - An **index** allows users to find all mentions of specific entities like people (e.g., Emil Sieg), places (e.g., Berlin), manuscript fragments, and specific Tocharian word forms across all the transcribed documents.
- The overall workflow is: **Transkribus** → **TEI** → **Enriched TEI** → **JSON** → **HTML** (for the website).

Notes from the presentation "Digital Humanities Meets Assyriology: A Look at Three Groundbreaking Projects in Vienna " by Nicla De Zorzi from University of Vienna.

Digital Humanities Meets Assyriology: Three Groundbreaking Projects in Vienna

This summary details three interconnected projects based in Vienna that apply digital humanities methods to the study of ancient Mesopotamian cuneiform texts.

---

1. Foundational Concepts: Cuneiform and the Akkadian Language

- **Cuneiform Script**:
    - The name comes from the Latin *cuneus*, meaning "wedge."
    - It is a writing system characterized by wedge-shaped impressions made with a stylus on soft clay tablets.
    - It was originally invented by the Sumerians to write their language and was later adapted for Akkadian.
- **Akkadian Language**:
    - Akkadian is an ancient Semitic language, part of the same family as modern Hebrew and Arabic.
    - It was the language of ancient Mesopotamia, encompassing the regions of Babylonia and Assyria.
    - The term "Akkadian" includes various dialects defined by region and time period.
    - **Standard Babylonian** was the high-prestige, literary version of the language, used across Western Asia from the 15th to the 4th century BCE. Mastering it required extensive training, typically available only to elite scribes in royal courts or temples.

Project 1: DigEanna & NaBuCCo – Digitizing the Eanna Temple Archive

This project focuses on making the vast administrative archive from the Eanna temple in Uruk accessible for historical research.

- **The Eanna Archive**:
    - **Source**: The archive originates from the Eanna temple, dedicated to the goddess Ištar in the southern Babylonian city of Uruk.
    - **Time Period**: It covers the late 7th to 6th century BCE, which is considered the best-documented period in Babylonian history.
    - **Scale**: It is a massive collection, comprising over **9,000 cuneiform tablets** and fragments. These are currently housed in various museums across the United States and Europe.
    - **Content**: The tablets are primarily administrative documents detailing the temple's economic and social activities.
- **Project Goals and Strategy**:
    - **Primary Goal**: To create a comprehensive digital analysis of the Eanna archive.

    - **"Digests over Editions"**: The project's key strategy is to prioritize creating **digests (paraphrases)** and extracting metadata (keywords, names) rather than producing full, line-by-line editions of every tablet.
    - **Efficiency**: This approach offers a practical tradeoff between speed and comprehensiveness, allowing researchers to quickly grasp the content of thousands of simple administrative texts.
    - **Selective Editions**: Full, detailed editions are reserved for only the most important and representative texts.

- **The Digital Workflow**:

    - **DigEanna & eBL**: Editions of selected tablets are created within the **DigEanna-Project** using the **Fragmentarium**, an open-source web application developed by the **Electronic Babylonian Literature (eBL)** project.
    - **NaBuCCo Platform**: The curated data—including metadata, digests, and full editions—is then exported via web services to the **NaBuCCo (A Neo-Babylonian Cuneiform Corpus)** platform, which serves as the main online catalogue.
    - **Interoperability and Export**: Data is available on both the eBL and NaBuCCo platforms for review and reuse. Both allow the export of these richly annotated texts in standard formats like JSON and **TEI XML**.

Project 2: Bestiarium Mesopotamicum – Unlocking Animal Omens

Led by Nicla De Zorzi and funded by the FWF, this project creates an open-access digital edition and a comprehensive analysis of animal omens from a major Babylonian divinatory text series.

A. The World of Mesopotamian Divination

- **The Role of Gods**: It was believed that the gods communicated with humanity by sending signs.
- **Two Forms of Communication**:
    1. **Provoked Divination**: A diviner could actively seek a sign by asking a question in a formal ritual context.
    2. **Unprovoked Divination**: The gods could send messages on their own initiative through everyday phenomena, like the sudden appearance of an animal, an earthquake, or an eclipse.
- **The World as a "Text"**: Diviners saw the world as a fabric into which the gods "wrote" messages. For example, the stars were called the "writing of the firmament" (*šiṭir burūmê*), and the liver of a sacrificial sheep was seen as a "tablet of the gods" (*ṭuppu ša ilī*).
- **The Diviners**: Divination experts were highly trained scribes who often served as royal advisors, using their skills as a decision-making tool for kings and elites.

B. The *Šumma ālu* ("If a City...") Omen Series

- **The Text**: The project focuses on a canonical Babylonian divinatory composition titled *Šumma ālu ina mēlê šakin* ("If a city is set on a height").
- **Massive Scale**: This series originally contained over **13,000 omens** organized into more than 100 thematic chapters.
- **Project Focus**: The *Bestiarium Mesopotamicum* project specifically targets the **47 chapters** that deal with animal behavior.

- **Animals Covered**: The omens feature a wide range of creatures: snakes, scorpions, insects, dogs, pigs, lions, foxes, birds of prey, and aquatic animals.

C. The Internal Logic and Structure of Omens

- **Fundamental Structure**: Omens are formulated as "if-then" statements:
    - **Protasis (Sign)**: The "if" clause that describes the observed event.
    - **Apodosis (Prediction)**: The "then" clause that gives the outcome.
- **Systematic Generation, Not Empirical Observation**:
    - These lists are not simple records of historical events.
    - They are highly schematic and systematically generated based on a set of intellectual principles.
- **Key Structuring Principles**:
    - **Binary Opposition**: Pairing omens based on opposites like up/down or right/left.
    - **Repetition with Variation**: Creating sequences where an omen repeats elements from the previous one but adds a new variable.
    - **Associative Lists**: Grouping omens by related concepts, such as sequencing predictions based on body parts, colors, or months of the year.
    - **Complex Literary Patterns**: The sequence of predictions (apodoses) can form elaborate structures, such as a **chiasm** or a circular pattern.
- **The Central Rule of Similitude**:
    - The link between the sign and the prediction is **always based on analogy or similarity** (semantic, phonetic, or graphic).
    - **Example**: "If a dog digs in the dirt in front of a man and lies down - his wife will be a serial adulteress." The dog's actions are seen as analogous to the wife's future behavior. This reflects a worldview where the universe is an interconnected web of similarities reflecting the divine will.

D. Case Study: The Birds in *Šumma ālu*

The study of birds in the omen series provides a clear example of how analogical thinking was applied. Birds were seen as significant due to a combination of unique characteristics:

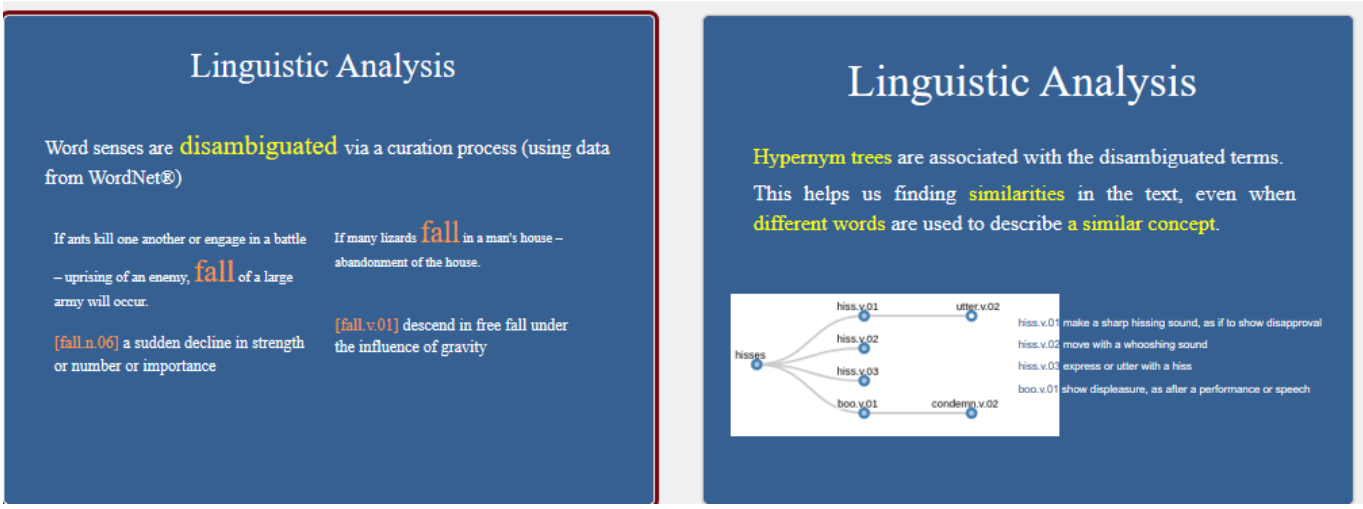
- Mobility in three dimensions.
- Curiosity and articulated sociality.
- Vocal abilities.

This perspective is famously summarized by Claude Lévi-Strauss, who noted that birds form an independent society that appears homologous to human society. The omen texts apply this analogy directly:

- **Positive Social Interaction**: "If a falcon and a raven eat something together – there will be peace that brings good fortune in the land".
- **Negative Social Interaction**: "If a falcon and an eagle don’t agree with each other and fight each other [...]" the prediction would logically relate to strife or conflict among humans.

E. The Digital Philology Workflow

- **The Challenge of Fragmentation**: The complete text of *Šumma ālu* must be reconstructed from hundreds of broken tablet fragments ("witnesses").
- **The "Score" as a Solution**:
  - The project digitally aligns the text from all parallel witnesses into a **score**.
  - From this score, a reliable **composite text** is created that reconstructs the full omen.
- **Custom Digital Tools and TEI Format**:
  - A custom-built web application processes Excel sheets containing the transliterations from each witness.
  - The final online edition is highly interactive, presenting the transliteration, transcription, and translation in parallel columns and tracking variant readings.
  - The final data can be exported in **TEI format**. The presentation shows a screenshot of the TEI XML file structure, demonstrating how individual witnesses, variant readings, and philological notes are encoded.
- **Advanced Linguistic Analysis**:
  - The project uses tools like **WordNet** to disambiguate the meaning of words.
  - Data visualizations like Sankey diagrams are used to explore the relationships between animals, actions, and the sentiment (positive, negative, uncertain) of the predictions.



Project 3: Comparative Divination Studies (Mesopotamia & Early China)

- This research extends the analysis to a comparative perspective.
- It explores structural parallels between the generative principles in Mesopotamian omen lists and those found in early Chinese divinatory texts written on bamboo strips, such as the **Shifa 筮法 (Stalk Divination Model)**.
- The core insight is that both cultures developed complex textual traditions that were not designed for simple linear reading but must be engaged with as a **"dynamic hypertext,"** tracing their intricate networks of internal correspondences and recursive loops.

Digital Approaches to Thomas Bernhard

Part 1: Introduction - Bernhard's View on a Technicized World

This section establishes the thematic foundation of the presentation by connecting Thomas Bernhard's literary concerns with the methods of digital analysis used to study his work.

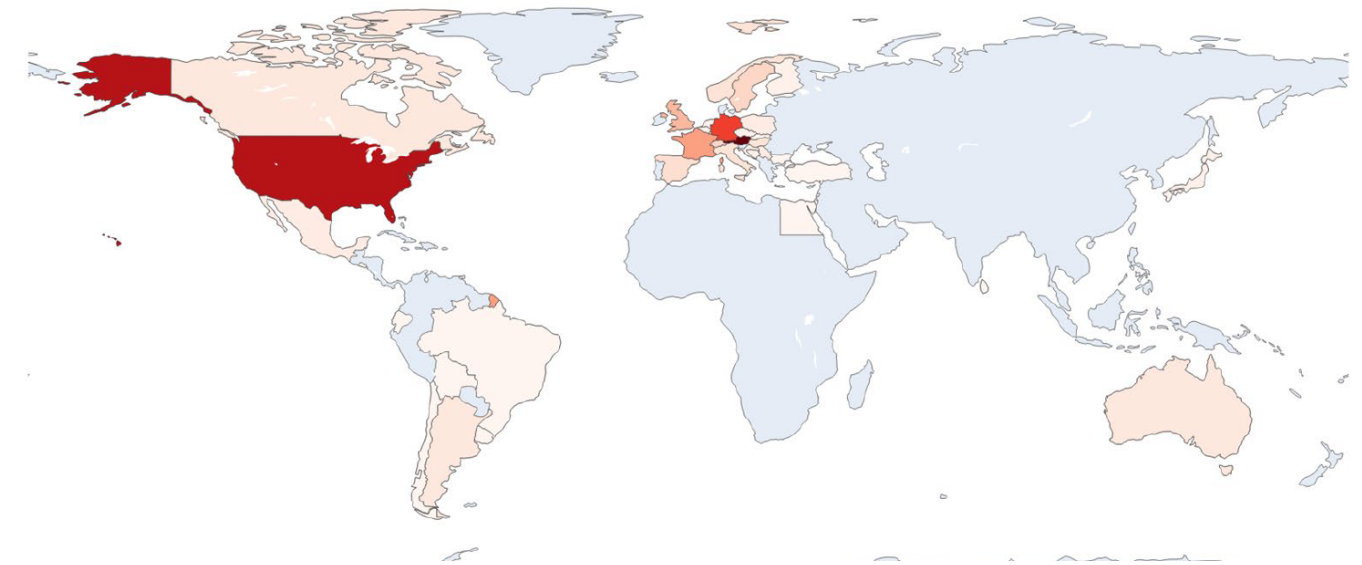
- **The World as a Machine:** Bernhard's writing is deeply preoccupied with the cold, rational, and "technicized" nature of the modern world. He saw humanity itself becoming mechanized.
    - **Direct Quote from *Verstörung* (1967):** "Calculating machines, that's all people are... The world is more and more just a computer."
    - **Obsessive "Studies":** This theme is prominent in novels like *Das Kalkwerk* and *Korrektur*, where protagonists are trapped in absurdly detailed, pseudo-scientific projects, reflecting a world where "Life is just science."
- **A Style of Deconstruction:** Bernhard's literary style mirrors his worldview. He advocated for dismantling holistic structures.
    - **Quote on Destruction:** "There must be nothing whole, one must smash it."
    - **Writing as a Destructive Game:** He described his own process as building up complex sentence structures like a child's toy, only to "smash everything together again."
- **The "Bernhard Machine":** His style became so iconic that it has been described by others as a predictable, repetitive "machine" or a "trick" (*Masche*). This recognizable, almost algorithmic style makes his work particularly suitable for computational analysis.

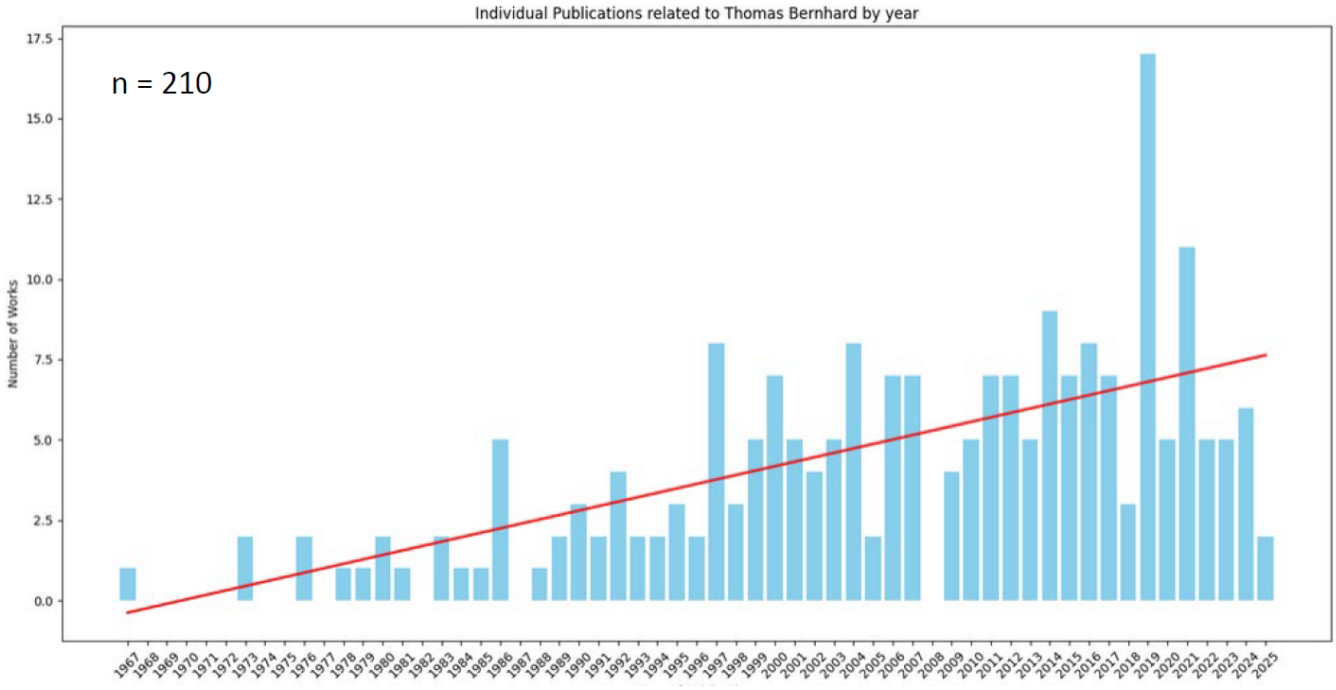
Part 2: Metadata & Databases - Mapping Bernhard's Universe

This section covers large-scale projects that collect and visualize data *about* Bernhard's work, its translations, and its influence.

Project 1: *Global Bernhard*

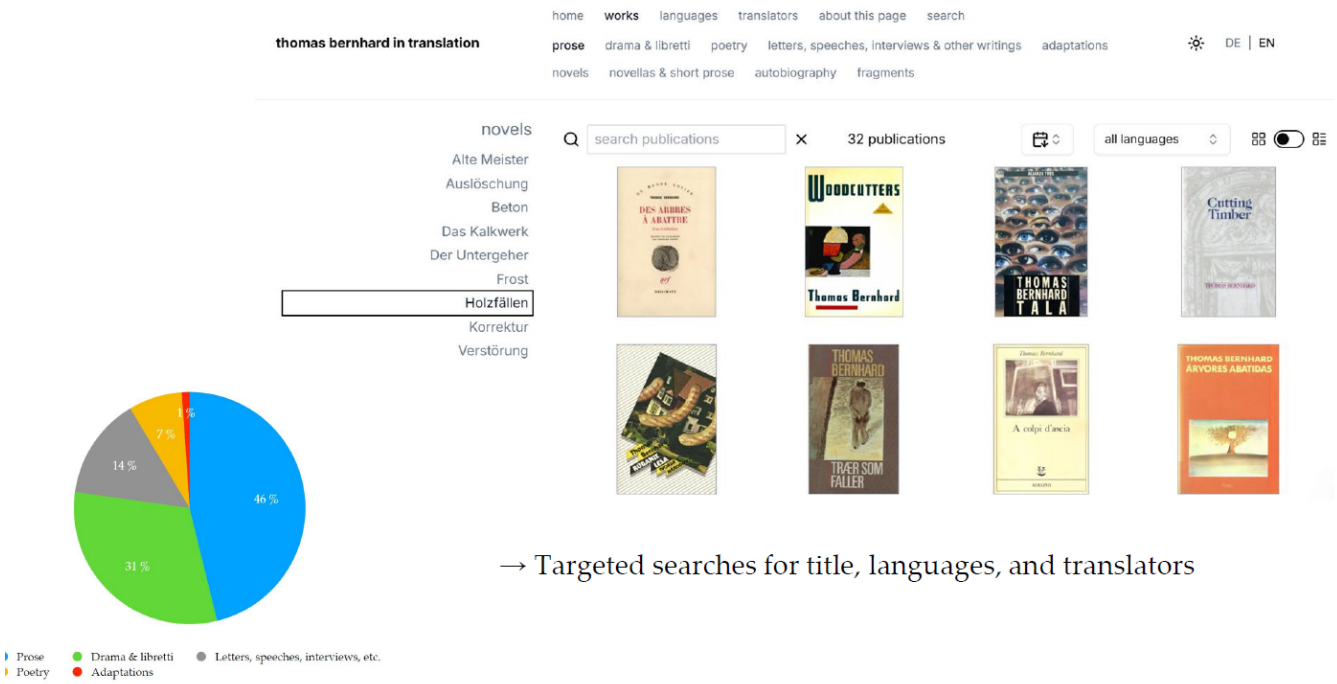
- **Purpose:** To track and document the international reception and influence of Thomas Bernhard on other writers. The project highlights how authors globally have engaged with, reacted to, or struggled against his powerful style, which has been called a "curse" and a "virus."
- **Methodology:** A database built on the TYPO3 platform collects examples of authors influenced by Bernhard, including bibliographic data and quotes.
- **Key Findings (from Data Visualizations):**
    - **Geographic Reach:** The strongest influence is seen in Austria and Germany, followed by the USA and France.
    - **Publication Trends:** There was a major spike in publications related to Bernhard around 1999, the 10th anniversary of his death. Publications about his reception and translations are increasing.
    - **Demographics:** The project allows for detailed analysis of author demographics, such as gender distribution across different generations.





Project 2: *thomas bernhard in translation*

- **Purpose:** The first comprehensive online database of all published translations of Bernhard's works.
- **Scope & Features:**
  - Contains over 1,000 entries for publications in 42 languages.
  - Users can perform targeted searches for specific titles, languages, or translators.
  - It provides rich metadata, including translators, publishers, and publication years.
  - **A unique feature** is the analysis of book covers, which reveals how marketing and reception are visually guided in different countries.
- **Key Findings:**
  - **Most Translated Works:** *Wittgensteins Neffe* and the autobiographical works (*Die Ursache*, etc.) are among the most frequently translated.
  - **Translation Peaks:** The highest number of first translations occurred in the 1980s and 1990s.



Future Project: *Forschungsstelle Thomas Bernhard*

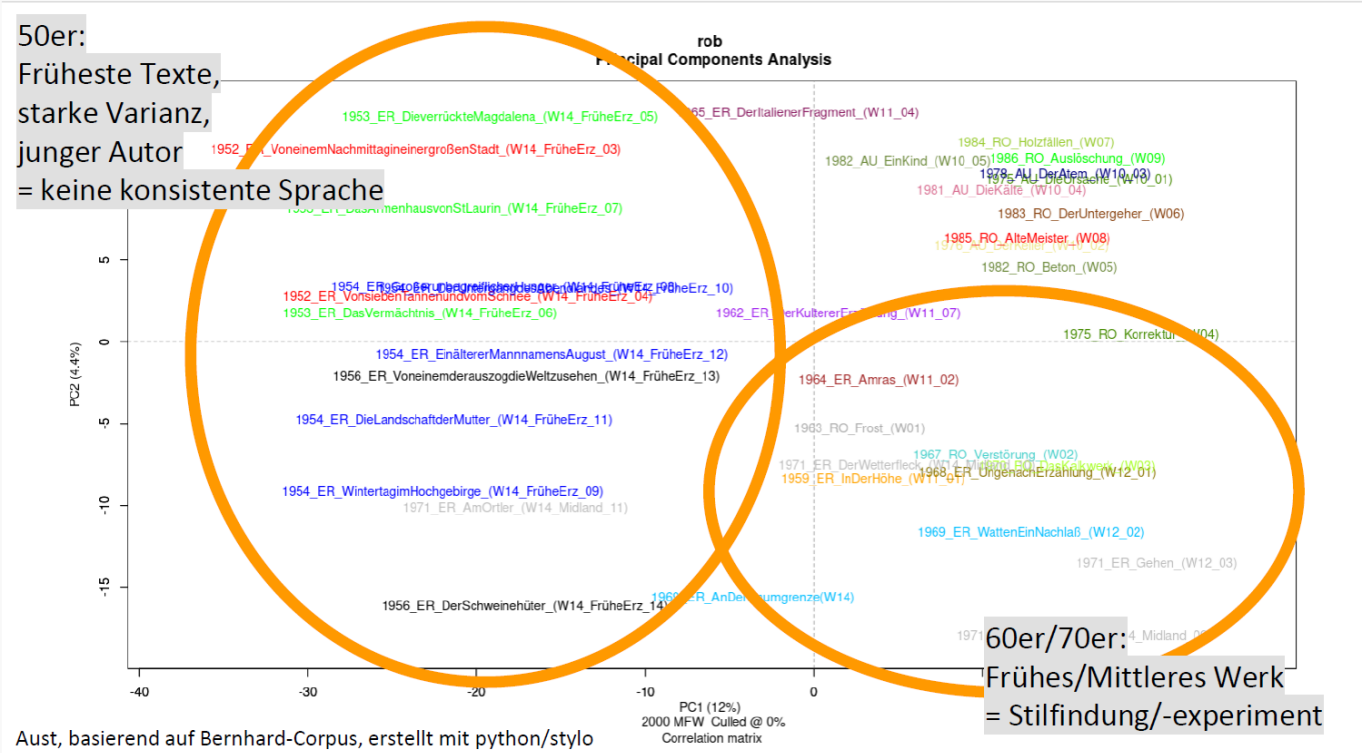
- This is a planned digital research hub that will integrate multiple databases (translations, theatre productions, reviews, etc.) into a single, interconnected platform using a flexible data model.

Part 3: Text Data & Text Mining - Deconstructing Bernhard's Style

This section applies computational methods directly to the text of Bernhard's prose to analyze his style with quantitative precision.

1. Stylometry: Mapping the Evolution of Bernhard's Style

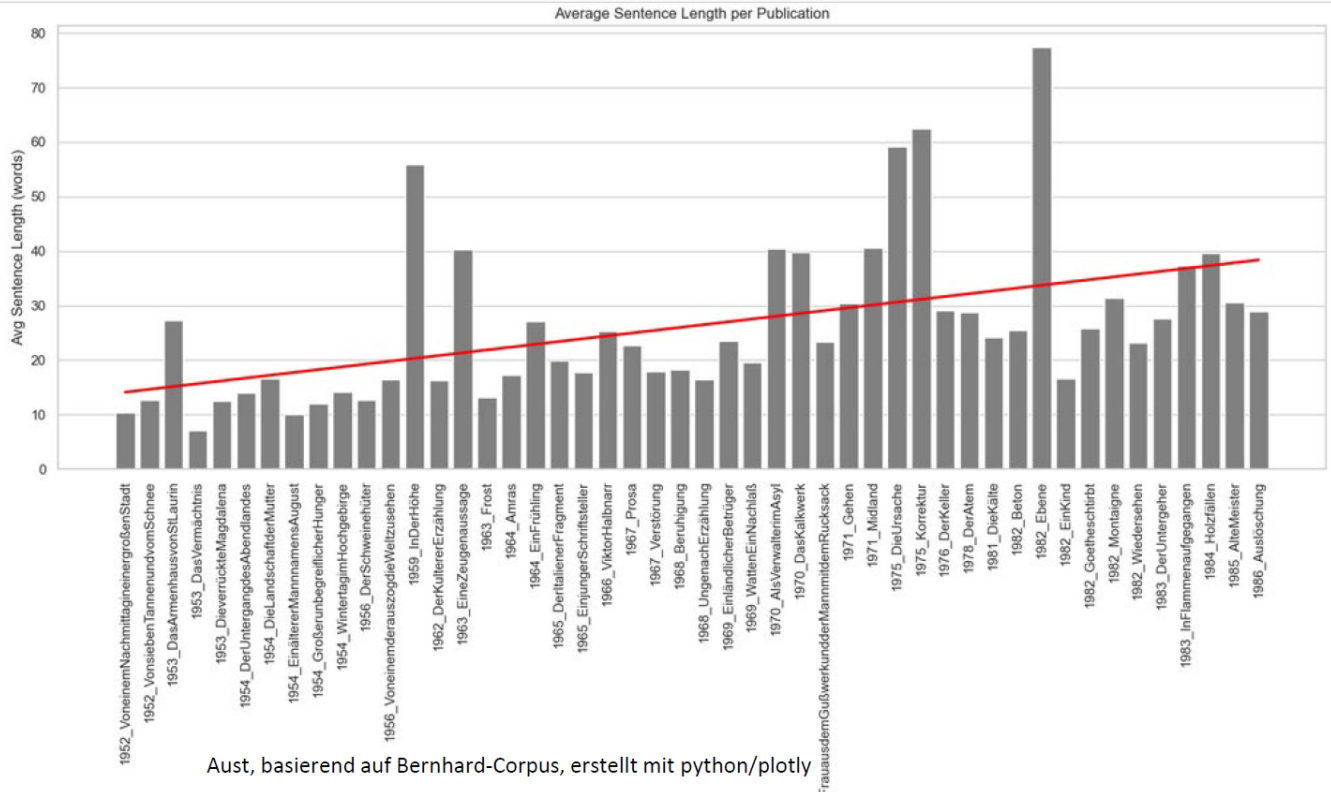
- **Methodology:** Stylometry measures stylistic similarity between texts by analyzing the frequency of the most common words (like "the," "and," "but"). It doesn't interpret meaning but reveals underlying authorial patterns. **Principal Components Analysis (PCA)** is used to visualize these similarities as clusters.
- **Key Findings:**
  - **Confirmation of Work Phases:** The analysis provides quantitative evidence for the traditionally defined phases of Bernhard's career.
  - **Early Work (1950s):** The texts are stylistically scattered, showing an author experimenting and finding his voice.
  - **Middle Work (1960s-70s):** The texts form a clearer cluster, demonstrating a period of stylistic consolidation.
  - **Late Work (1980s):** The late novels (*Holzfällen*, *Alte Meister*, *Auslöschung*) are clustered extremely tightly, proving that he had developed a highly stable and recognizable late style.
  - **Autobiography:** His five autobiographical books form their own distinct stylistic group, separate from his other fiction.



Aust, basierend auf Bernhard-Corpus, erstellt mit python/stylo

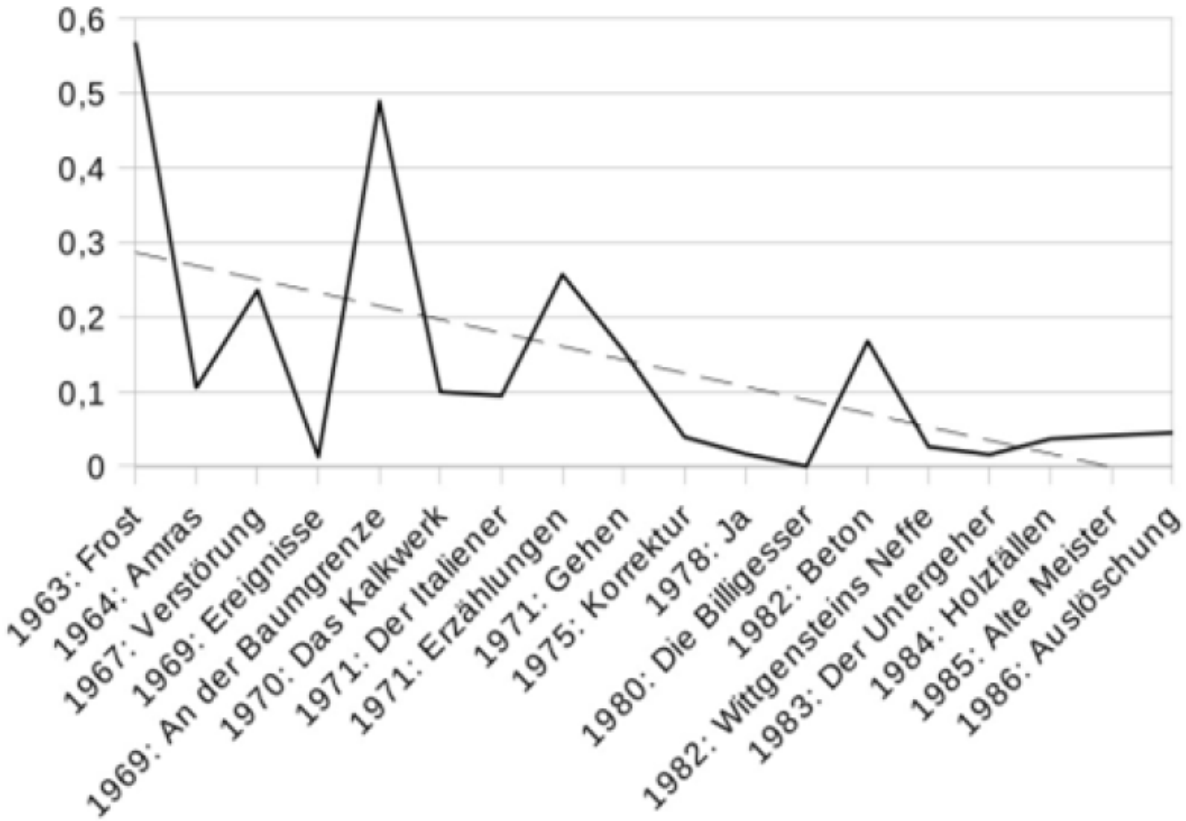
2. Linguistic Repetition and Sentence Structure

- **Sentence Length:**
  - There is a clear trend of increasing average sentence length over his career.
  - The analysis pinpoints extreme examples, such as the single sentence in *Korrektur* that is **1,156 words long**. Visualizing sentence lengths in his novels like *Korrektur* shows a pattern of "swelling and receding" waves, which has been compared to musical structures like Ravel's *Bolero*.



Aust, basierend auf Bernhard-Corpus, erstellt mit python/plotly

Abb. 13: Prozentualer Anteil (Y-Achse) von Fragezeichen in ausgewählten Bernhard-Prosatexten (X-Achse)

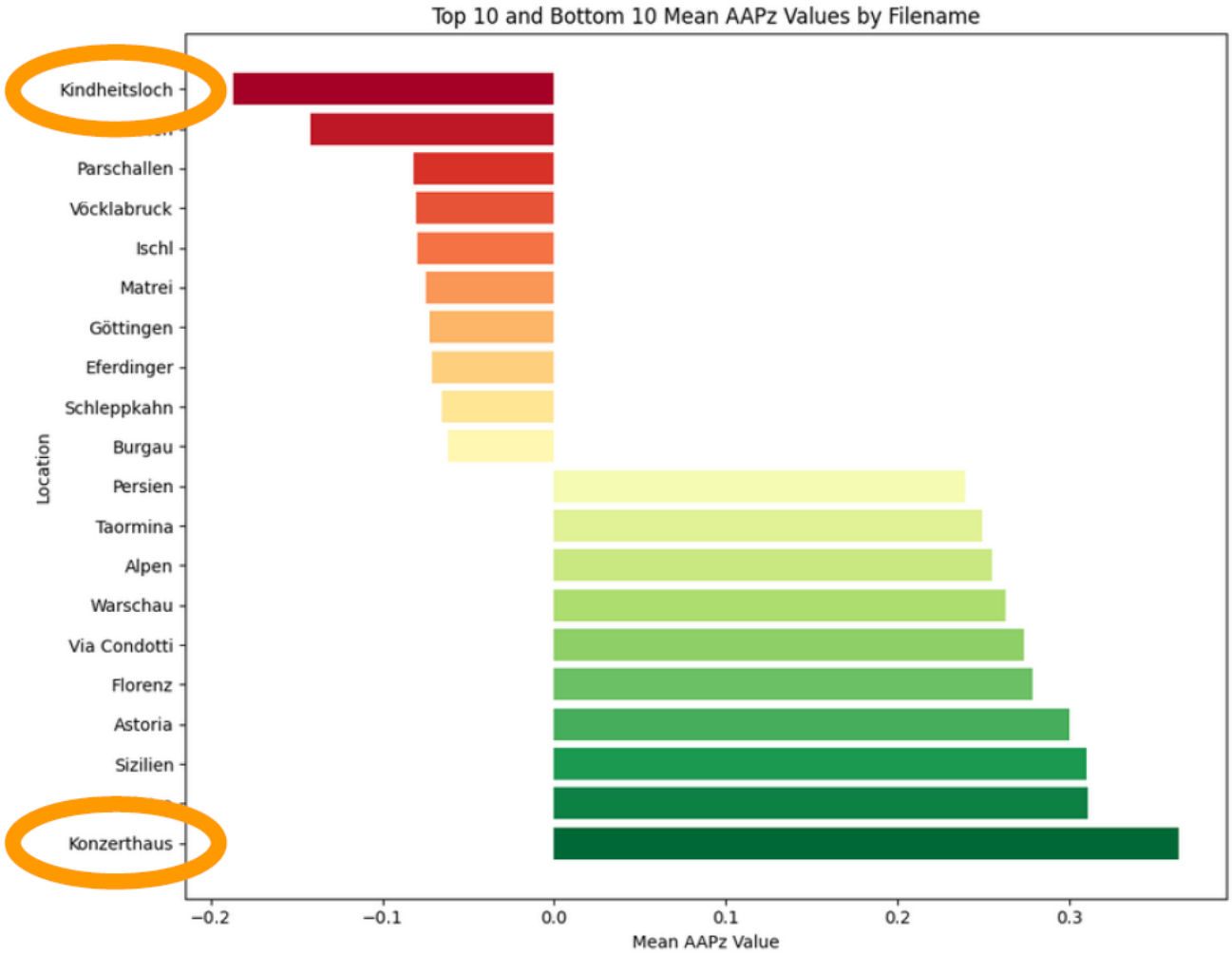


- **Lexical Diversity (Type-Token Ratio - TTR):**
  - The TTR, which measures vocabulary richness, shows a consistent **decrease** over his career. This quantitatively proves that his language became more repetitive and lexically sparse over time.
- **"Bernhardisms":** The analysis tracks the frequency of his signature words and phrases (e.g., *naturgemäß*, *sogenannt*, *dachte ich*, *wie ich sagen muss*). The usage of these "tics" varies and helps define his different work phases.

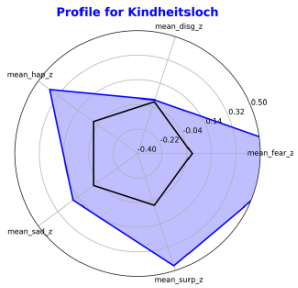
3. Sentiment Analysis: Mapping Emotions in Places

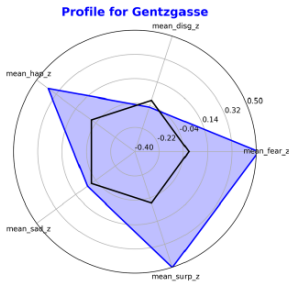
- **Methodology:** This innovative approach combines two techniques:
  1. **Named Entity Recognition (NER):** Automatically identifies all place names (e.g., "Vienna," "Rome") in the texts.
  2. **Sentiment Analysis:** Uses a lexicon (SentiArt) to calculate the average emotional "valence" (positive/negative polarity, anger, fear, happiness, etc.) of the sentences in which each place name appears.
- **Key Findings:**
  - **Beyond "City-Insults":** While Bernhard is famous for his rants against Austrian cities, the analysis reveals a more complex emotional landscape.
  - **The Positive South:** Locations in Southern Europe (Sicily, Florence, Rome) are consistently associated with a more positive emotional context, fitting their role as places of escape in his later works.

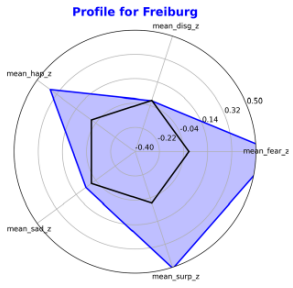
- **Psychological vs. Real Places:** The most intensely emotional locations are not actual cities but **"psychological spaces."**
  - Most Negative: ***Kindheitsloch*** ("childhood hole"), a term for the traumatic space of his youth.
  - Most Positive: ***Konzerthaus*** ("concert hall"), representing the transcendent power of art.
- **Emotion Profiles:** This method can generate unique "emotion profiles" for different locations, visualizing the specific mix of sadness, happiness, fear, etc., associated with each one.

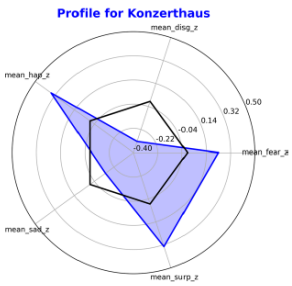


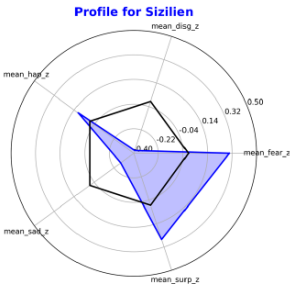
‘Emotionsprofile’

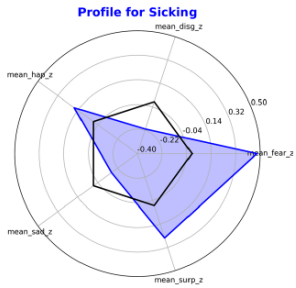












Schubert-digital

---

1. Introduction to Digital Musicology

Digital Musicology is a sub-field of Digital Humanities that applies computational methods to musical sources and data.

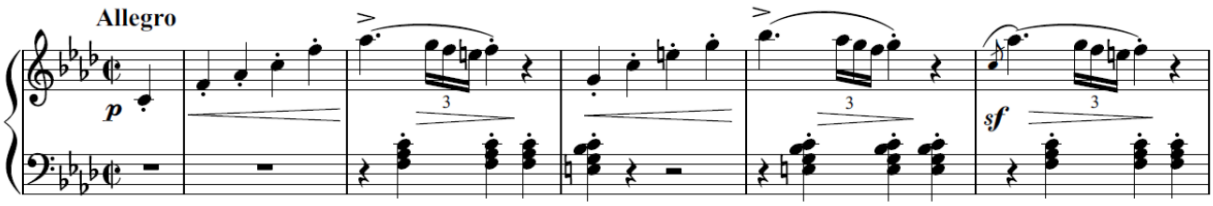
- **The Core Challenge of Music:** Music is fundamentally a "time-ordered set of sounds" (for the listener) or "gestures" (for the performer). Creating a static, readable representation (like a score) is a complex translation that involves many assumptions and conventions.
- **The "Big Tent" of Digital Musicology:** Like Digital Humanities, it's a broad field encompassing many activities:
  - **Music Corpus Studies:** Creating and analyzing large digital collections, like catalogues of a composer's complete works.
  - **Digital Music Philology:** The creation of scholarly digital editions of musical texts.
  - **Music Performance Analysis:** Studying recordings and performance data.
  - **Computational Music Theory:** Using algorithms to analyze musical structures.

---

2. Music Encoding: Turning Scores into Data

Music encoding is the process of representing musical notation in a structured, machine-readable format. This is essential for any computational analysis.

- **Complexity:** It's more complex than encoding plain text because a score contains multiple layers of information:
  - **Score Properties:** Key signatures, time signatures, number of staves.
  - **Note Properties:** Pitch (e.g., C, F#), octave, duration (e.g., quarter note), articulations (staccato, accent), stem direction, beams, and slurs.



score definition
properties:
@ number of staves
@ brace
@ key / accidentals
@ meter

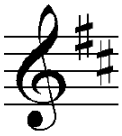
staff definition
properties:
@ lines
@ clef: on which line?
[@ key / accidentals
@ meter]

<note>
properties:
@name
@octave
@duration
@accidentals
@stem-direction
@dots
@articulation

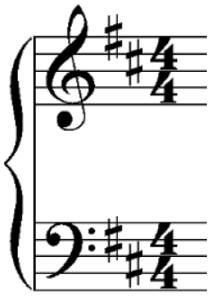
- **History of Encoding Formats:**

- The first format was **DARMS** (1960s).
- Other formats include Humdrum, MuseData, and the widely used **MusicXML**.

- **The Music Encoding Initiative (MEI):**
  - MEI is the current gold standard for scholarly work. It is an **XML-based system** modeled on the Text Encoding Initiative (TEI).
  - Its goal is to create a highly detailed, machine-readable structure for encoding music documents, flexible enough for everything from simple songs to complex orchestral scores.
  - MEI captures not just the notes, but also information about the physical source, editorial changes, and different versions.
  - A tool like **Verovio** is needed to render the MEI code visually as a standard musical score.



```
<staffDef clef.line="2" clef.shape="G" key.mode= "major" key.sig="2s" lines="5" n="1"/>
```



```
<scoreDef meter.count= "4" meter.unit="4" key.sig="2s">
        <staffGrp n="1" symbol="brace" label="Pianoforte" >
                <staffDef clef.line="2" clef.shape="G" lines="5" n="1"/>
                <staffDef clef.line="4" clef.shape="F" lines="5" n="2"/>
        </staffGrp>
</scoreDef>
```

MEI captures not just the notes, but also information about the physical source, editorial changes, and different versions. Application scenarios of the MEI header: Scholarly editions of music: Indication of different sources and scribes' hands are needed for the critical apparatus. Work catalogue: Giving information about the work without using the music section of the file. Performance database: Specification of the performance data is also possible in the header.

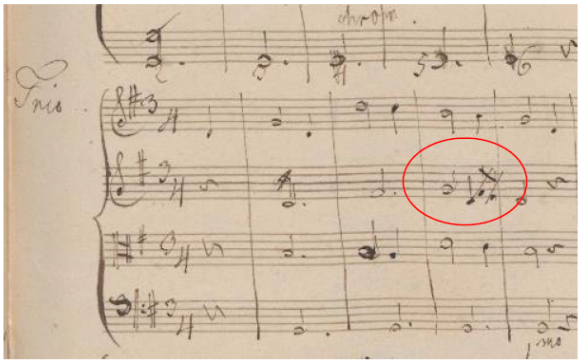
---

3. Scholarly Digital Editions

The goal of a scholarly edition is to publish a musical text based on a critical and thorough examination of the historical sources (e.g., manuscripts, early prints). MEI is the ideal tool for this.

- **The 9 Steps of Editorial Work:**
  1. Source research
  2. Deciphering (reading the handwriting)
  3. Source description
  4. Source evaluation (determining which is most reliable)
  5. Comparing readings (different versions)
  6. Determining source dependency (e.g., this copy was made from that original)
  7. "Higher" criticism (making editorial judgments)
  8. Producing the musical text
  9. Producing the critical report (explaining all decisions)

- **Encoding Editorial Changes:** MEI allows editors to digitally tag specific actions found in the manuscript, such as deletions (`<del>`), additions (`<add>`), or substitutions (`<subst>`), and attribute them to a specific person (e.g., the composer or a later editor).



A-Wn, Mus.Hs. 44706. pag. 73.

```
<staff n="2">
  <layer n="1">
    <note pname="g" oct="4" dur="2"/>
      <subst hand="#Anton_Bruckner1">
        <del rend="strike">
          <beam>
            <note pname="f" oct="4" dur="8" accid.ges="s"/>
            <note pname="e" oct="4" dur="8"/>
          </beam>
        </del>
        <add>
          <note pname="d" oct="4" dur="4"/>
        </add>
      </subst>
  </layer>
</staff>
```

---

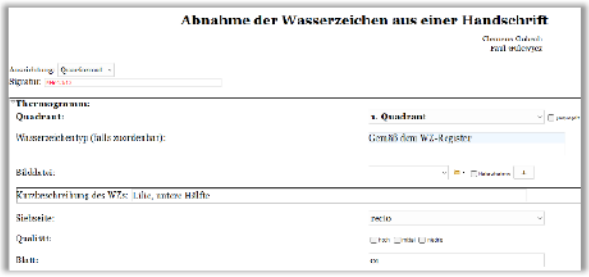
4. The *Schubert-digital* Project

Schubert-digital is an online research platform applying these digital methods to the autograph (handwritten) manuscripts of Franz Schubert.

- **Scope:** The project aims to document the roughly **500 surviving Schubert autographs** (totaling ~4,700 pages), most of which are held in Viennese collections.
- **Primary Aims:**
  1. **Detailed Source Description:** To create a comprehensive digital record of every manuscript's physical characteristics.
  2. **Digital Reunification:** To digitally reassemble manuscripts that were broken apart and scattered across different libraries after Schubert's death.
- **The Data Model:** The project uses a sophisticated data model to distinguish between:
  - `Gesamtmanuskript`**:** The complete, reconstructed manuscript as it was originally created by Schubert.
  - `Bestandteil`**:** The physical fragment or section of the manuscript as it exists today in a library.
- **Reconstructing Paper Structure:** MEI is used to encode the physical structure of the manuscript gatherings (quires). It records how individual sheets of paper (`bifolium`) were folded and nested together. This allows for a virtual reconstruction of the original manuscript booklet.

```
<watermark label="a">
    <title>Gemäß dem WZ-Register</title>
    <fig>
        <graphic/>
        <figDesc>
            <heraldry>
                <annot type="description">
                    <p>Lilie, untere Hälfte</p
                </annot>
            </heraldry>
            <term label="Siebseite">recto</ter
            <term label="Qualität"/>
        </figDesc>
    </fig>
```

```
<locus label="Blatt">01</locus>
<locusGrp label="Doppelblatt">
    <locus>01</locus>
    <locus>08</locus>
</locusGrp>
<locusGrp label="Papierbogen">
    <locus>02</locus>
    <locus>07</locus>
</locusGrp>
<term label="Format"/>
<annot type="comment">
    <p/>
</annot>
</watermark>
```



---

5. Watermark Digitization & Thermography

Watermarks are designs embedded in paper that are crucial for dating a manuscript and tracing its paper to a specific mill.

- **How Watermarks Are Made:** In manual papermaking, a wire design (the watermark) is sewn onto the wire mesh of a mould. This mould is dipped in paper pulp, and the paper ends up slightly thinner where the wire design was.
- **The Challenge:** Watermarks can be very faint and are often obscured by the ink of the musical notation written on top of them.
- **Solution: Thermography**
    1. A manuscript page is placed on a gently heated copper plate.
    2. A high-resolution **infrared camera** takes a picture of the page.
    3. Heat travels faster through the thinner parts of the paper (the watermark).
    4. The resulting image, a **thermogram**, clearly reveals the watermark design, free from the visual interference of the ink.

---

6. Advanced Analysis: Digital Reconstruction & Automated Clustering

The thermographic images of watermarks and paper structure enable powerful new research methods.

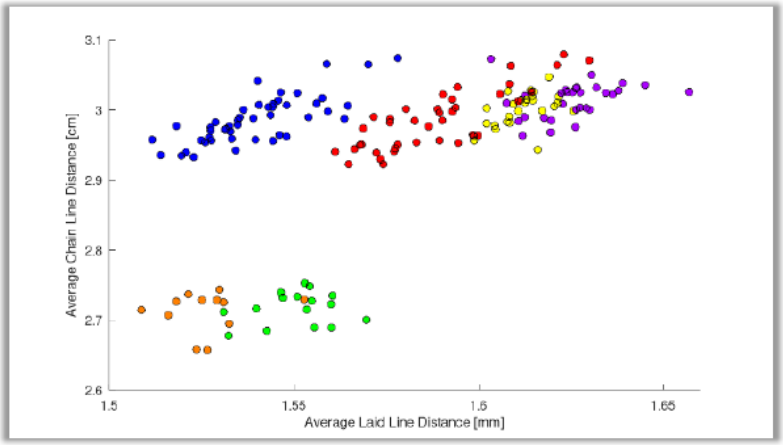
- **Digital Reconstruction:** By identifying matching watermarks on scattered pages, researchers can digitally "stitch" them back together to reconstruct the original, full paper sheet from which they were cut.

[Image Placeholder: Diagram showing two separate manuscript pages being digitally combined to form one original sheet, Slide 102]

- **Automated Clustering (The Paper's "Fingerprint"):** The project uses algorithms to analyze the physical structure of the paper itself, which is unique to the mould that made it. This allows for grouping papers even without a clear watermark.
    - **Method 1: Radon Transformation:** Detects and measures the precise distance between the vertical **"chain lines"** in the paper's mesh.
    - **Method 2: Laid Line Density:** Measures the density of the very fine horizontal **"laid lines."**
- **The Result:** By plotting these two measurements on a graph, papers made on the same physical mould form tight, distinct **clusters**. This technique is so precise it can distinguish between "twin" moulds—nearly identical moulds that were used in pairs to speed up paper production. This provides an objective, data-driven way to group manuscripts, helping to refine their dating and better understand Schubert's working process.

Automated Clustering

- Example: Welhartitz, Quadrant 4, both twins
- Type I-1: red/blue
- Type I-4: yellow/violet
- Type I-6: green/orange



Quantum Computing Fundamentals & Applications in the Humanities

This summary synthesizes two presentations: "Quantum Computing" by Prof. Frank Leymann and "From Quantum Computing to Quantum Humanities" by Dr. Johanna Barzen, delivered at the University of Vienna in May 2025. It covers the fundamental principles of quantum computing and demonstrates a practical application in the digital humanities.

Part 1: Fundamentals of Quantum Computing

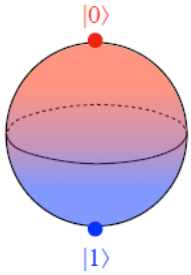
The Strange New World of Quantum Physics

The quantum world operates on principles that are fundamentally different from our everyday classical world. While the classical world is deterministic, the quantum realm is inherently **stochastic**, or probabilistic. You can only know the probability of a measurement's outcome, not the certain result beforehand. This leads to counter-intuitive phenomena.

- **Superposition:** A quantum system can exist in a combination of multiple states simultaneously. The most famous illustration of this is the **Schrödinger's Cat** thought experiment, where a cat in a box is considered both alive and dead until the moment it is observed.
- **The Qubit:** The classical bit is the basic unit of information, existing as either a 0 or a 1. The quantum equivalent, the **qubit**, can be a 0, a 1, or a superposition of both at the same time. Mathematically, this is represented on the **Bloch Sphere**, a unit sphere where any point on the surface represents a possible state for the qubit. This gives a single qubit access to an infinite number of potential states before it is measured.

Bit

A bit is either "0" or "1"
→ Two possible values



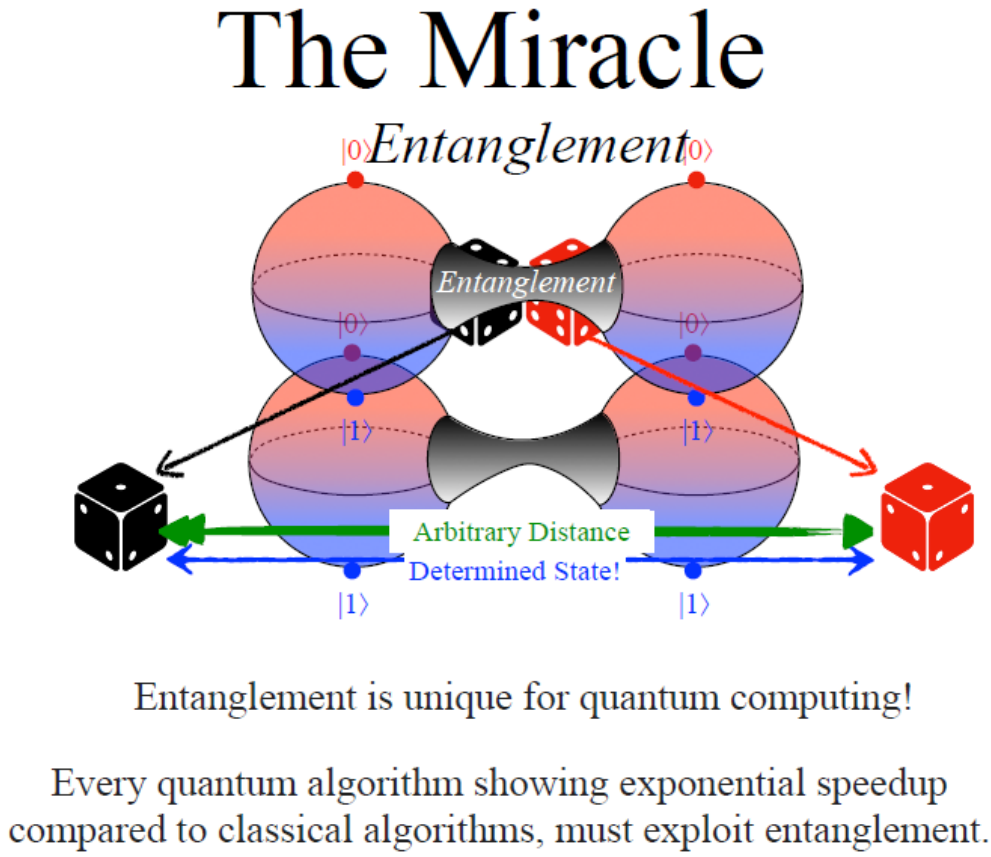
Qubit

A qubit is an arbitrary point
on the Bloch sphere
→ Infinitely many possible values

The Power of Quantum Computing

- **Quantum Parallelism:** The true power emerges when qubits are combined. A quantum register with 'n' qubits can exist in a superposition of all 2^n possible classical states at once. Applying a single operation to this register manipulates all 2^n values simultaneously. A system of just 300 qubits could process more values in an instant than there are atoms in the known universe.

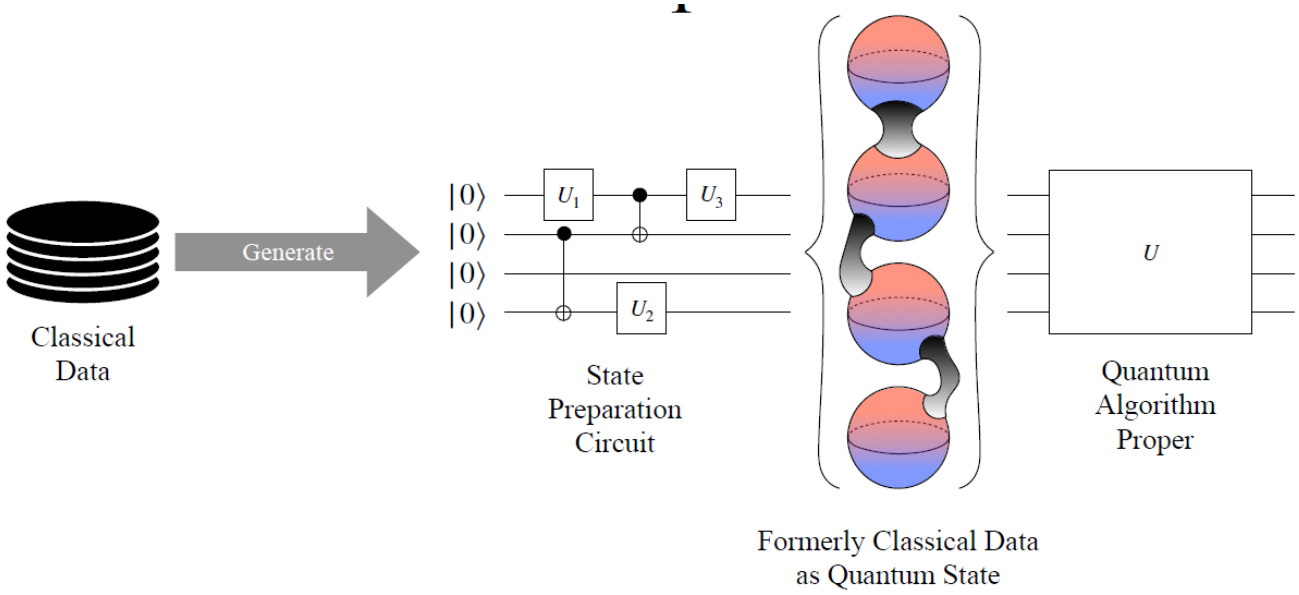
- **Entanglement:** Described by Einstein as "spooky action at a distance," entanglement is a unique quantum connection between two or more qubits. Once entangled, the state of one qubit instantly influences the other, regardless of the physical distance separating them. This non-local connection is a critical resource; any quantum algorithm showing an exponential speedup over its classical counterpart must leverage entanglement.

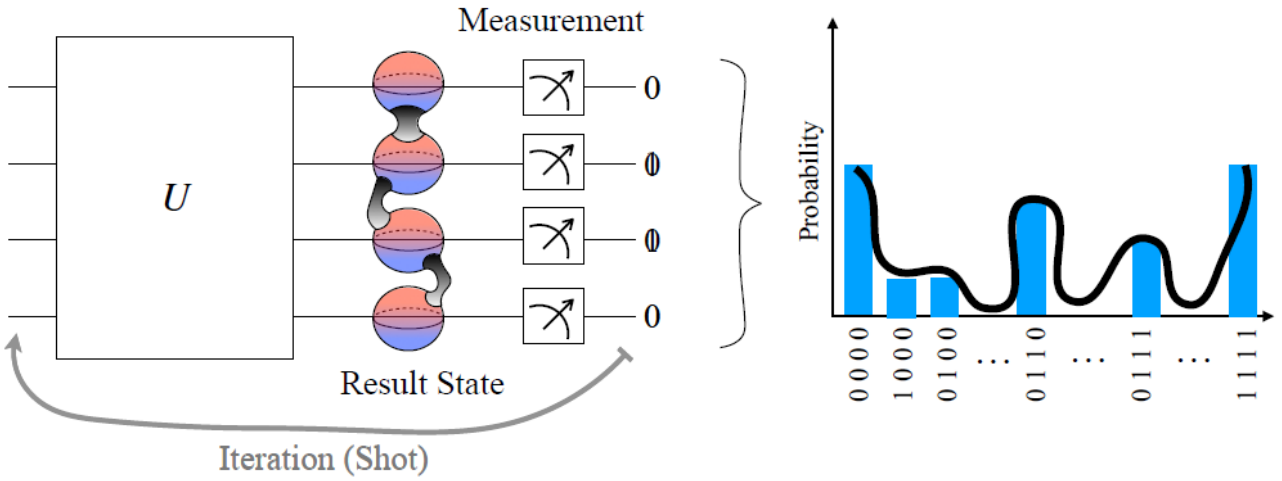


How Quantum Algorithms Work

Most quantum algorithms today are **hybrid**, using both a classical computer (CPU) and a Quantum Processing Unit (QPU).

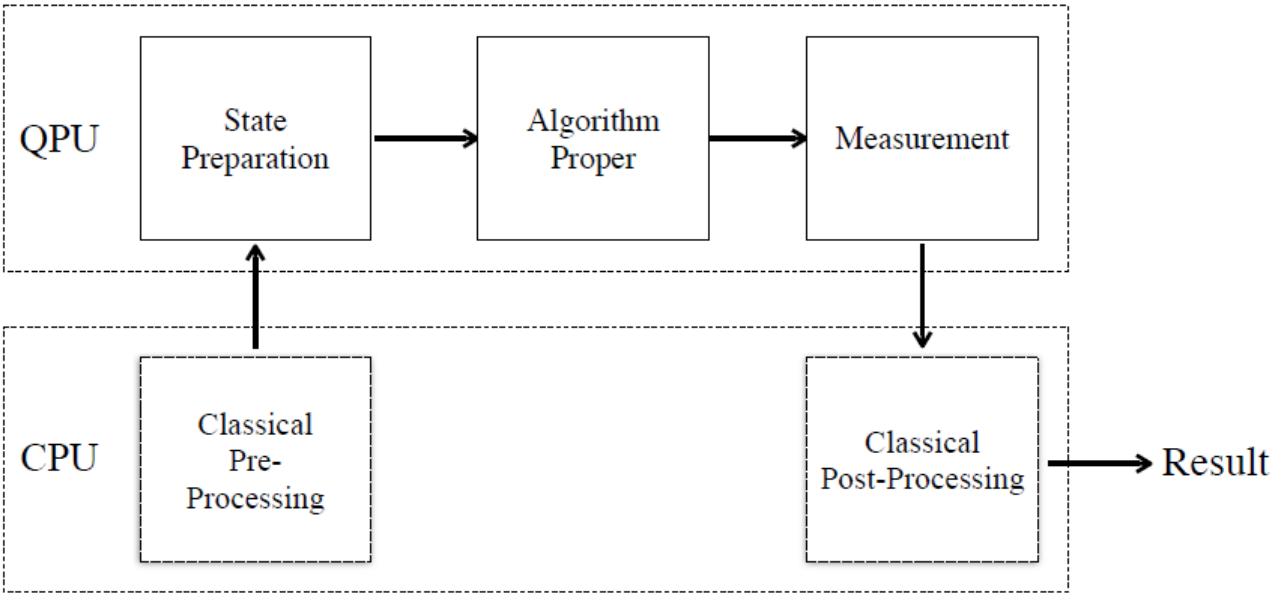
1. **Classical Pre-Processing:** A CPU prepares the initial data.
2. **State Preparation:** The data is loaded into the QPU, encoding it as a quantum state.
3. **Quantum Computation:** A sequence of operations, represented as a **quantum circuit**, is executed on the qubits.
4. **Measurement:** The qubits are measured, which collapses their superposition into a definite classical state (0s and 1s).
5. **Classical Post-Processing:** Because measurement is probabilistic, the computation is repeated thousands of times ("shots"). The CPU analyzes the resulting probability distribution to determine the final, most likely answer.





The result of a quantum computation is a probability distribution (most often)

This distribution is post-processed to get the "classical" result



Quantum Algorithms are hybrid (most often)

Training Quantum Neural Nets

- Classical No-Free-Lunch theorem of supervised learning

  The more training data is used,
  the lower the average error in learning a neural net

- Quantum No-Free-Lunch theorem of supervised learning

  The more the training data is entangled,
  the less training data is needed
  to learn a quantum neural net with low average error

  A ***single*** pair of maximally entangled training data suffice,
  to train a quantum neural net with low average error
  ("in high dimensions")

Applications and The Cryptographic Threat

Quantum computers promise exponential speedups for specific, complex problems.

- **Key Applications:**
  - **Factoring (Shor's Algorithm):** Can break today's standard encryption (like RSA) by efficiently finding the prime factors of large numbers.

    - **Unstructured Search (Grover's Algorithm):** Provides a quadratic speedup for searching databases, weakening symmetric encryption.
    - **Simulation:** Can accurately simulate molecules for drug discovery and material science.
    - **Machine Learning & Optimization:** Can solve complex optimization problems and enhance machine learning algorithms like Support Vector Machines (SVMs).

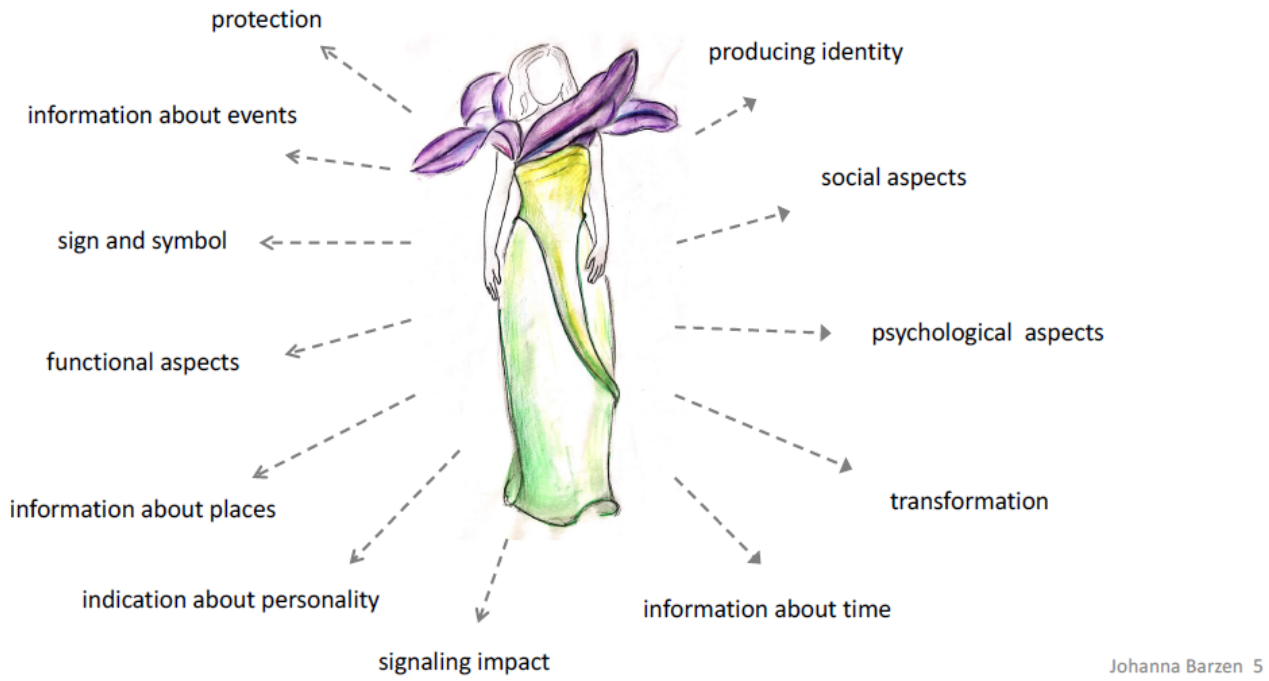
- **The Threat and The Solution:** The power of Shor's algorithm poses an imminent threat to global cybersecurity. In response, researchers have developed **Post-Quantum Cryptography (PQC)**—new cryptographic standards (like Kyber and Dilithium) designed to be secure against attacks from both classical and quantum computers.

Part 2: From Quantum Computing to Quantum Humanities

(Based on the presentation by Dr. Johanna Barzen)

Use Case: Analyzing Clothing in Film with MUSE

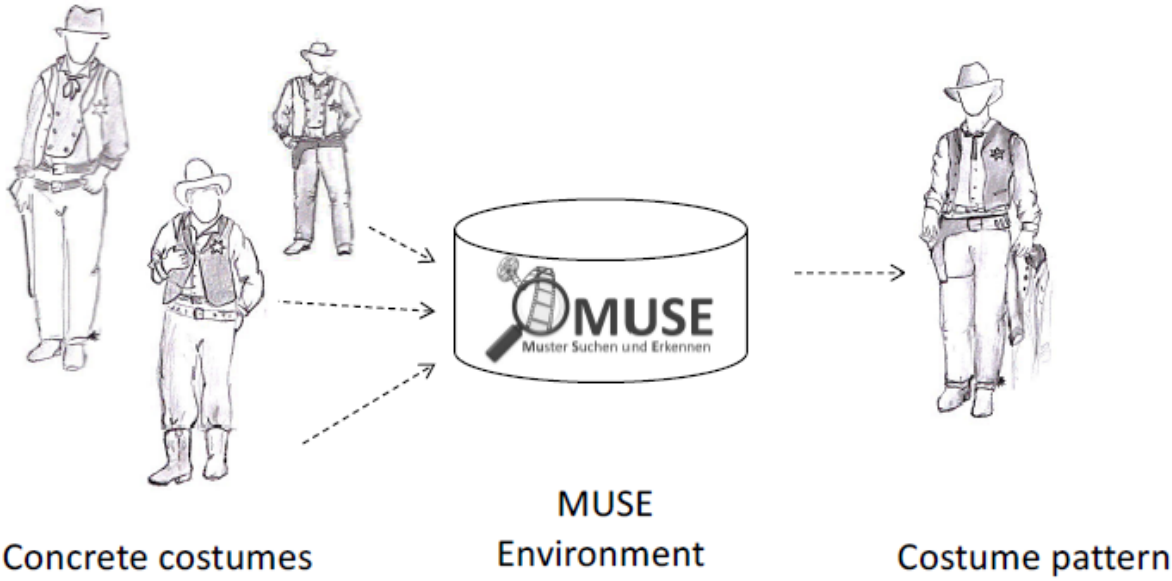
This section demonstrates how quantum computing can be applied to a humanities problem: analyzing "vestimentary communication," or the stories told by clothing in film. The **MUSE** project was created to develop a formal language for understanding costumes.



The core concept is that recurring character types or narrative situations are often represented by recurring **costume patterns**. The MUSE project built a massive repository to identify these patterns.

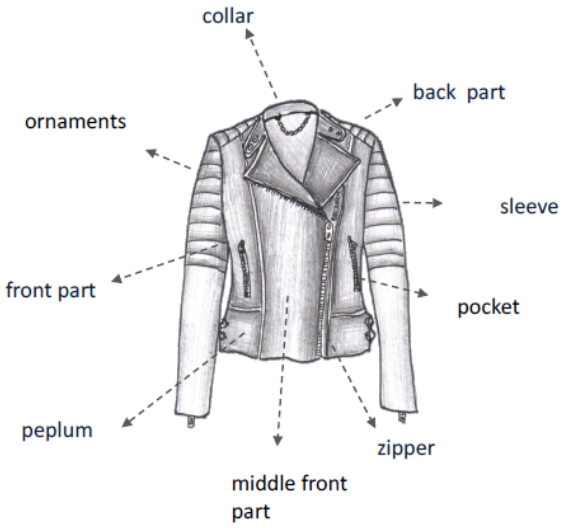
- **The MUSE Process:**
    1. **Define a Corpus:** Select genres with strong costume conventions (e.g., Westerns, High School Comedies).
    2. **Create a Repository:** Meticulously catalog every costume from the selected films using a highly detailed **taxonomy** (with 910 nodes) and **ontology** (with over 3,000 nodes) that break down garments into their base elements, materials, colors, shapes, etc.

3. **Identify Patterns:** Analyze the data to find proven solutions to recurring design problems. For example, the **"Male Outlaw"** pattern in Westerns is consistently communicated with a combination of 'long pants', 'shirt', 'revolver', 'cowboy boots', 'Akubra' hat, and 'gilet'.

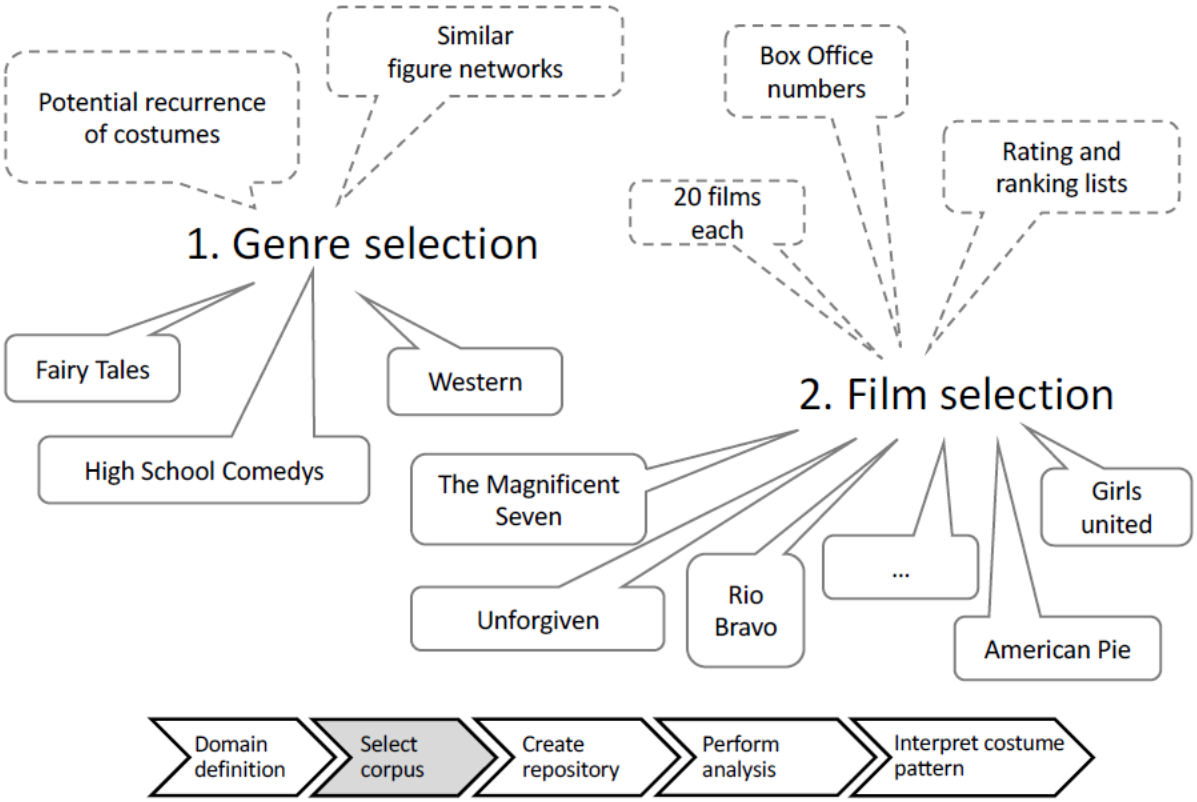


Costume-Relevant Parameters

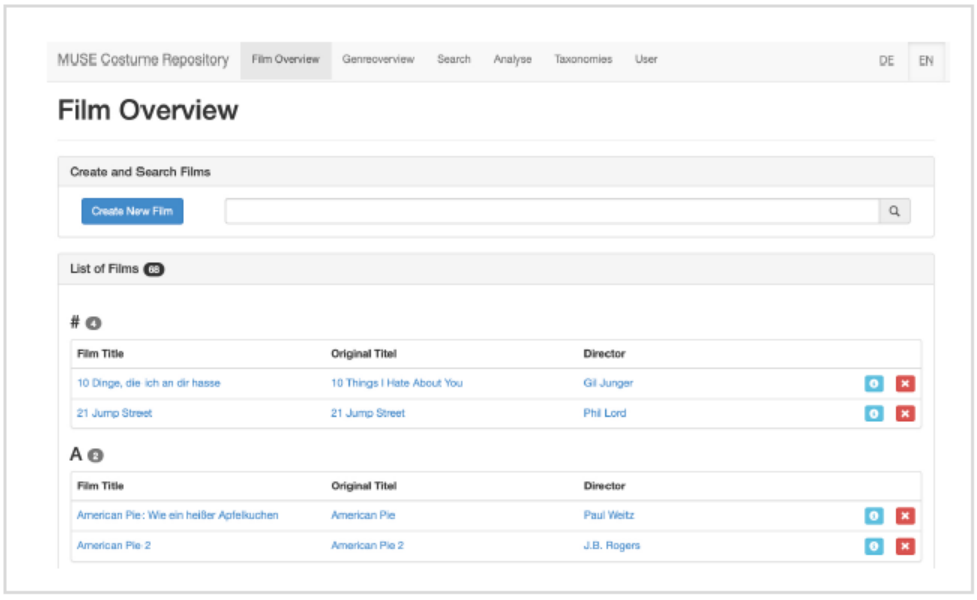
- Base elements
- Primitives
- Designs
- Shapes
- Material
- Colour
- Status
- Ways of wearing
- Function
- Body modification



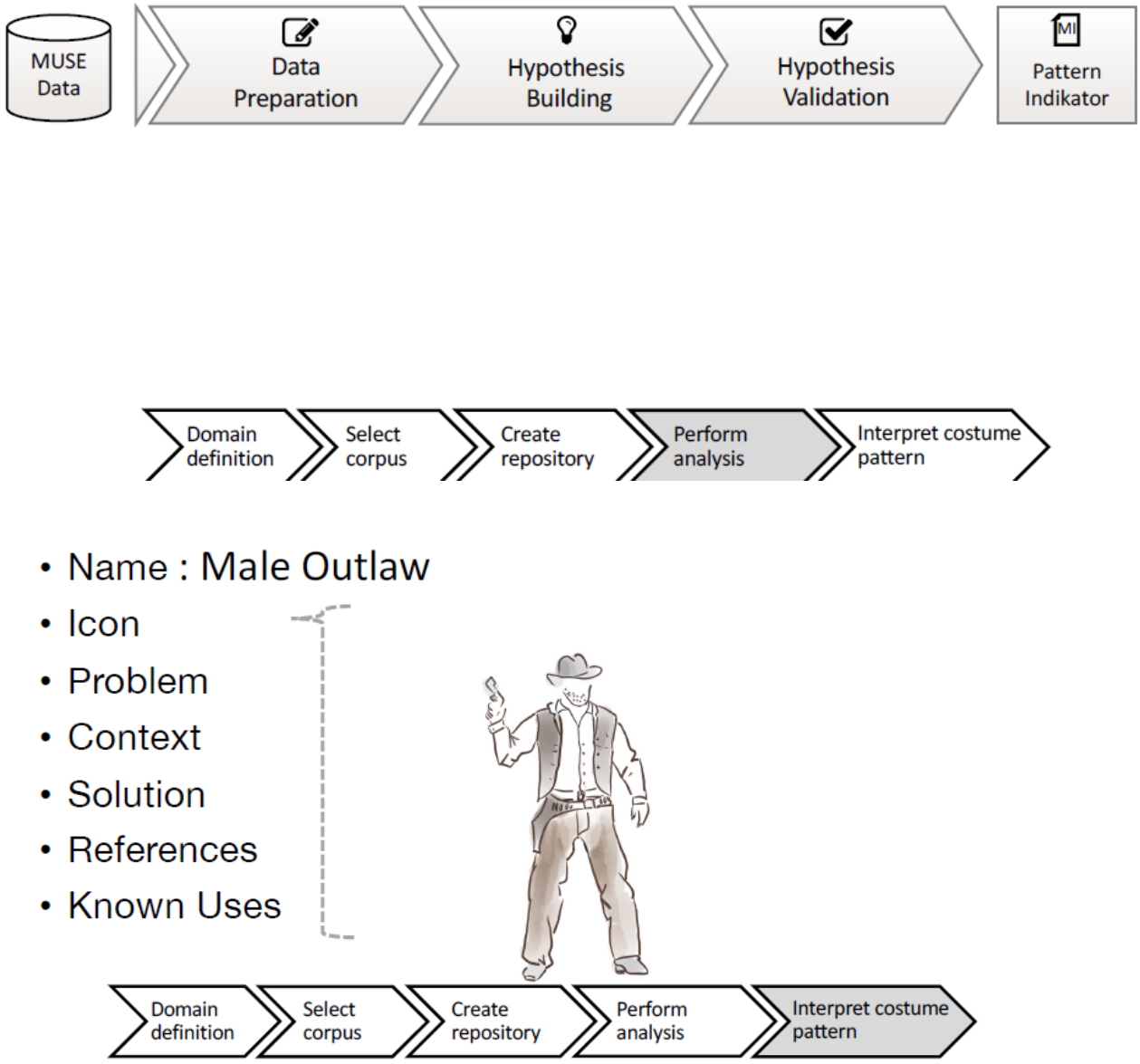
Johanna Barzen 23



Johanna Barzen 27



Johanna Barzen 28

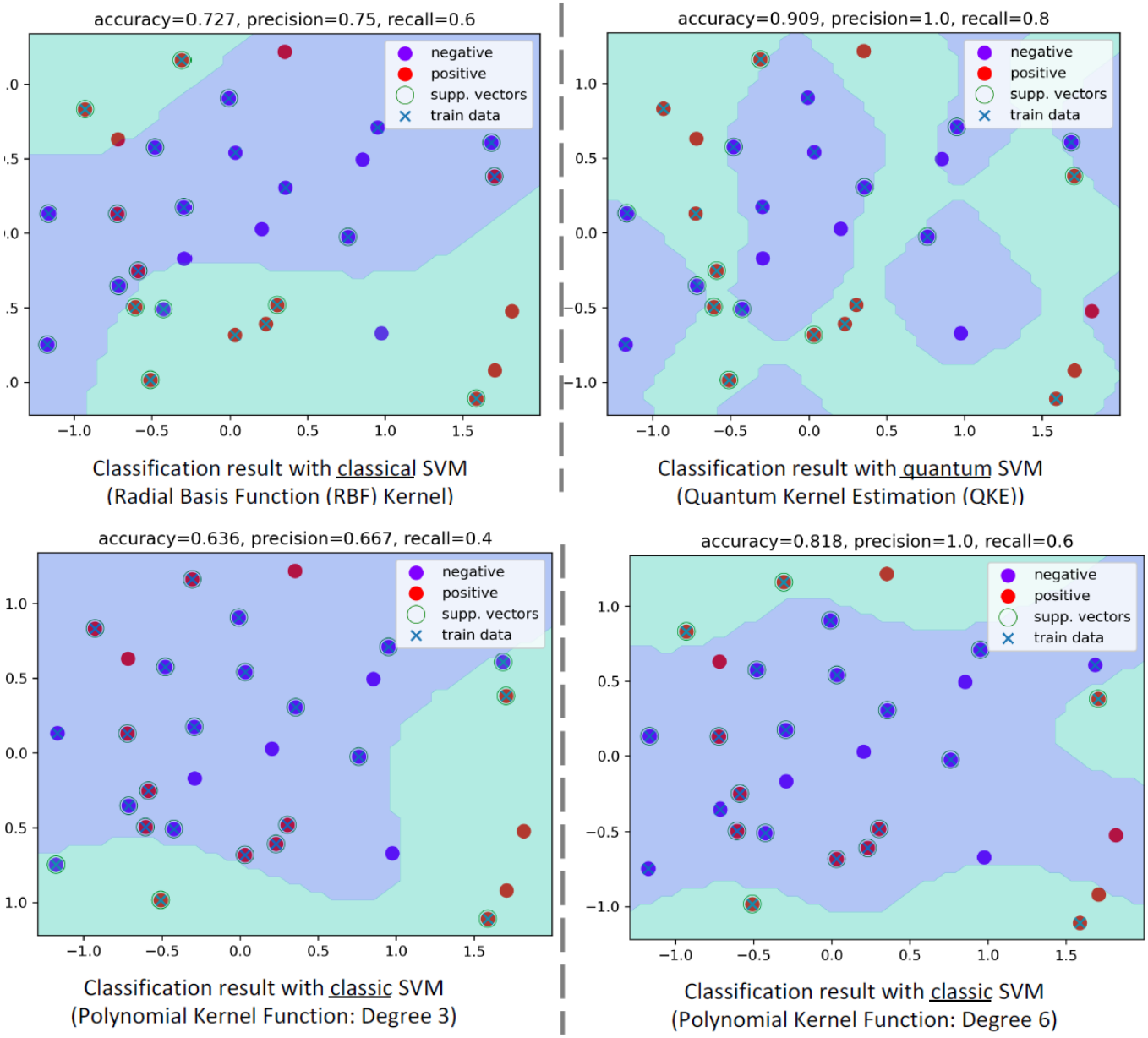


Introducing Quantum Humanities with QHAna

The massive, complex dataset produced by MUSE is an ideal candidate for quantum analysis. The **QHAna (Quantum Humanities Analysis Tool)** was developed to bridge the gap between humanities research and quantum computing. It allows non-experts to apply and compare classical and quantum machine learning algorithms.

- **The Challenge of Categorical Data:** Humanities data is often categorical ('red', 'denim', 'worn') not numerical. QHAna solves this by using the **Wu-Palmer Similarity** method. This algorithm calculates a numerical distance between two items in the MUSE taxonomy based on their positions in the hierarchical tree, transforming the qualitative data into a quantitative distance matrix that algorithms can process.

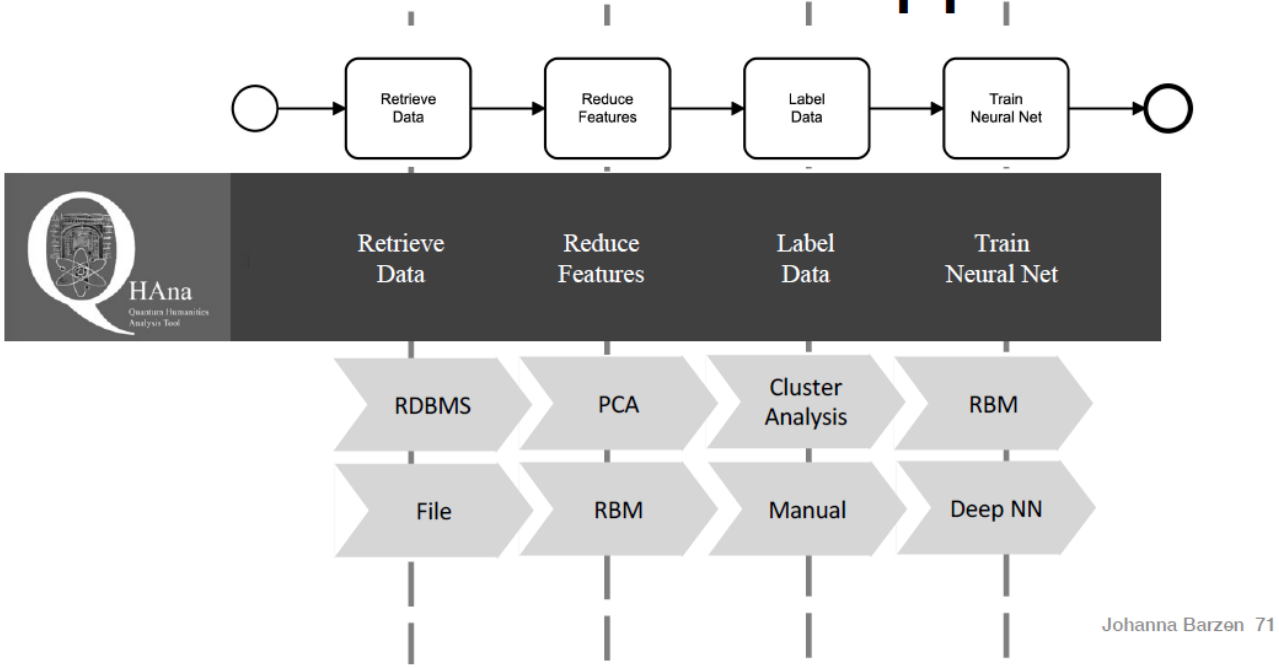
- **Comparing Classical and Quantum Results:** QHAna was used to perform clustering and classification on the MUSE costume data. The results showed that quantum algorithms could outperform their classical counterparts. For instance, the **quantum SVM** (using Quantum Kernel Estimation) achieved a significantly higher classification accuracy (90.9%) than the classical SVM (72.7%).



Goals and Outlook

QHAna is an open-source, plugin-based tool designed to be independent of any single use case. Its goal is to significantly ease the process of applying advanced computational methods to humanities research. By providing easy access to both classical and quantum tools, it empowers researchers to explore complex cultural datasets in entirely new ways, marking a concrete step from Digital Humanities toward **Quantum Humanities**.

Reengineering of QHana: Individual Workflow Support



Johanna Barzen 71

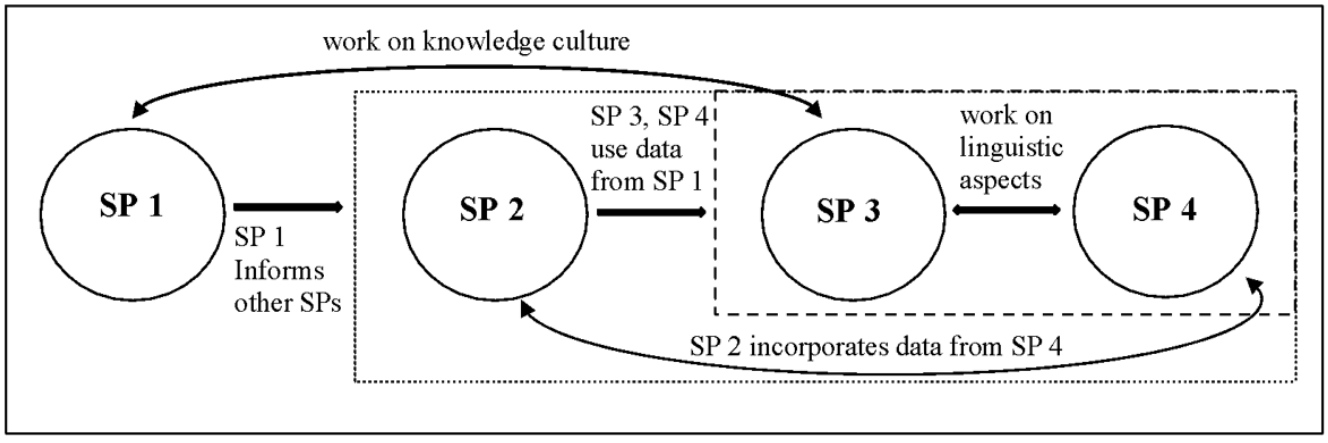
ARITHMETIC Project: An Overview

The **ARITHMETIC** project is a major academic study funded by the European Research Council (ERC), involving research teams from the Universities of Innsbruck and Graz.

- **Core Focus:** The project investigates German arithmetical treatises found in manuscripts from the late Middle Ages, specifically the period of 1400-1522.
- **Central Research Question:** How did the practice of arithmetic develop and spread in the German vernacular (the common language) during the transition from the Middle Ages to the Modern Period?
- **Significance:** While the history of Latin mathematical theory is well-documented, little is known about how practical math skills reached new groups of people, like merchants, through texts written in German. This project fills that gap, exploring the intersection of philology, history, and culture through a digital edition of these important texts.

Project Structure & Goals

The project is structured into four interconnected subprojects (SPs), each with a specific focus, that work together to build a comprehensive understanding of late medieval arithmetic.



Subproject 1: Documentation and Context

- **Goal:** To study the manuscripts as physical objects to understand their production, circulation, and use. This involves analyzing not just the text but the entire context.
- **Methods:**
    - **Codicology:** A detailed description of each manuscript's content, images, history, and a paleographical assessment.
    - **Comparative & Structural Codicology:** Comparing manuscripts to get information on book production and analyzing a codex as a whole to understand its compilation.
    - **Digital Analysis:** Using network analysis and visualization to map the distribution of texts and define manuscript groups.

Subproject 2: Transcription, Encoding, and Annotation

- **Goal:** To create a semantically enriched digital edition of the manuscripts. This goes beyond a simple transcription to enable deep, interdisciplinary research.
- **Methods:**
    - **"Assertive Edition":** A method that combines traditional textual criticism with digital methods of annotation and knowledge formalization.
    - **Ontology Building:** Conceptualizing an ontology (a formal model of knowledge) for late medieval arithmetic to make connections of knowledge visible within the discourse.

Subproject 3: Philological and Historical Analysis

- **Goal:** To analyze the content of the texts to draw conclusions about the social and collective knowledge of their users and writers.
- **Methods:**
    - **Discourse Analysis:** Using linguistic and historic-literary methods like lexeme-analysis and analysis of metaphors.
    - **Historical Pragmatism:** Analyzing the text to understand the communicative situations and the influence of oral traditions.
    - **Cultural Connection:** Identifying how other discourses of the time (e.g., religion, trade, other sciences) influenced these arithmetic texts.

Subproject 4: Development of Arithmetical Jargon

- **Goal:** To study the emergence of a specialized German language for arithmetic, paying close attention to the bilingual (German/Latin) environment of the time.
- **Methods:**
    - **Lexicographical Process:** Compiling a comprehensive glossary of arithmetic terminology. . This involves identifying terms (lemmas), tracing them back to their Latin origins if possible, and analyzing how their meanings may have shifted.
    - **Analysis of Neologisms:** Studying the creation of new German terms that did not have a direct Latin template.

The Digital Edition: Process and Technology

The core of the project is the creation of a digital edition, which relies heavily on modern technology for transcription and analysis.

The Research Corpus

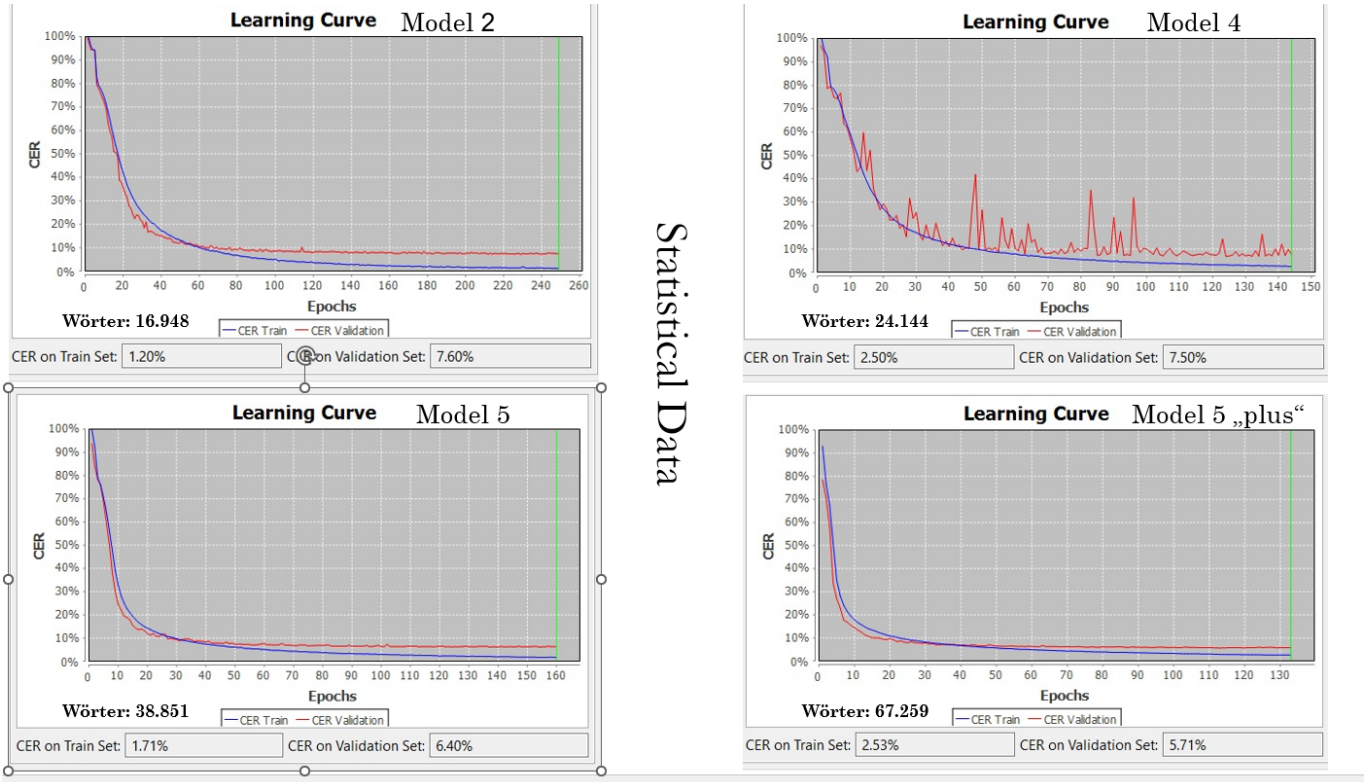
- The project works with a total of 135 manuscripts, with 116 being directly relevant.
- The primary focus is on the **60-70 manuscripts** created before or around the year 1500.
- As of May 2025, transcriptions have been started for most of the corpus, with the pre-1500 manuscripts nearly complete.

[Screenshot: An image of a manuscript from the corpus, for example, from the Kremsmünster Monastery Library. (from slide 35)]

Handwritten Text Recognition (HTR) with Transkribus

The project uses the **Transkribus** software to automate the transcription of these difficult handwritten texts.

- **Model Training:** HTR models are "trained" by providing them with manually transcribed text ("Ground Truth"). The project's process was iterative:
  - **Early Models (Model 2, 4):** Started with a small number of manuscripts. The Character Error Rate (CER) was initially high but improved as more data was added.
  - **The Breakthrough (Model 5):** A significant improvement was achieved by training the model on a larger corpus of manuscripts from the 15th century with similar handwriting styles.
  - **Generic Model:** The team created a powerful generic model for 15th-century texts, trained on **28 manuscripts** and over **286,000 words**.
- **Performance:** The models' accuracy steadily increased. The CER (a measure of errors) on the validation set dropped from **7.60%** with Model 2 to **5.71%** with the "Model 5 plus".



- **HTR Challenges:** Even with advanced models, challenges remain, especially with:

- Complex fractions.
- Multi-line calculations and tables.
- Pages with unusual or non-linear layouts.

[Screenshot: An image of a challenging manuscript page with a "mindmap" layout, such as ÖNB, Cod. 3502. (from slide 70)]

The Annotation Process

Annotation adds layers of meaning to the raw text. This is a two-step process:

1. **In Transkribus:** Structural elements are tagged, including headers, initials, glosses, and a detailed list of mathematical figure types (e.g., `multiplication_table`, `regula_de_tri`, `interest_calculation`).
2. **In an XML Editor (Oxygen):** Deeper semantic annotation is performed to tag key entities like **People**, **Places** (trade cities, routes), **Goods**, and **Currencies**. The mathematical content of entire paragraphs is also categorized.

The Final Product & Future Work

Features of the Digital Edition

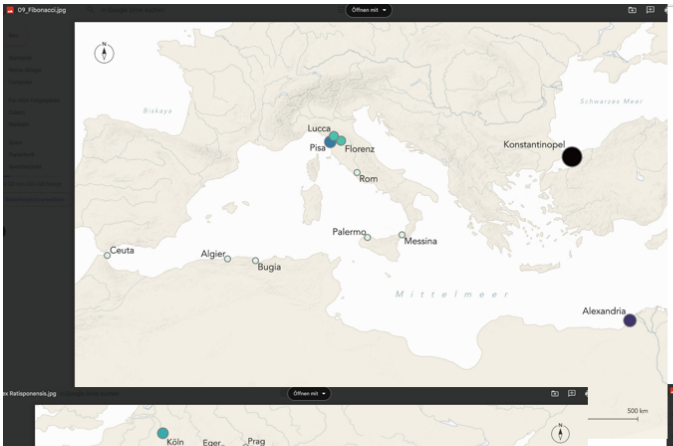
The publicly accessible digital edition will provide:

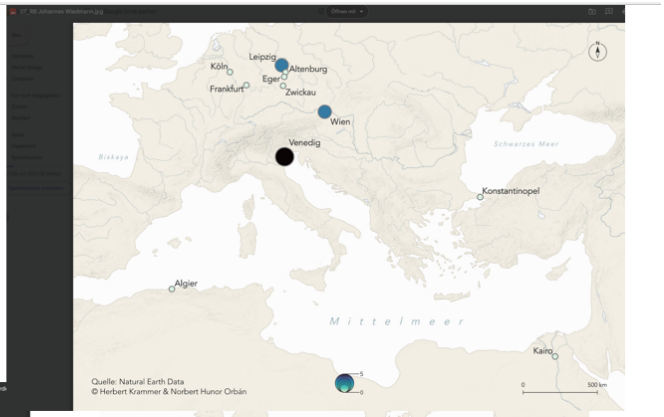
- High-resolution images of each manuscript.
- Both a diplomatic (as-is) and an expanded transcription.
- A downloadable, comprehensive glossary of mathematical and mercantile terms, complete with commentary and links to Latin sources.
- Detailed metadata for each manuscript.

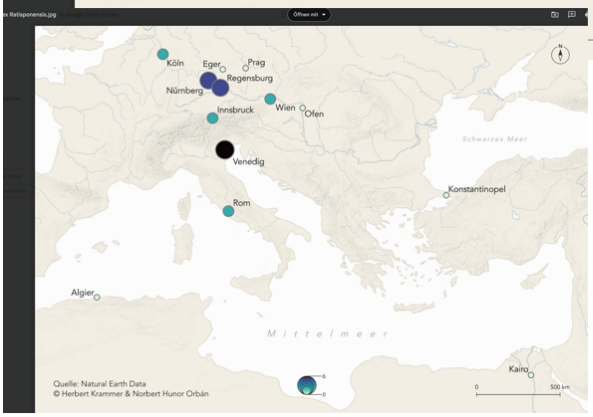
Future Analysis and Research

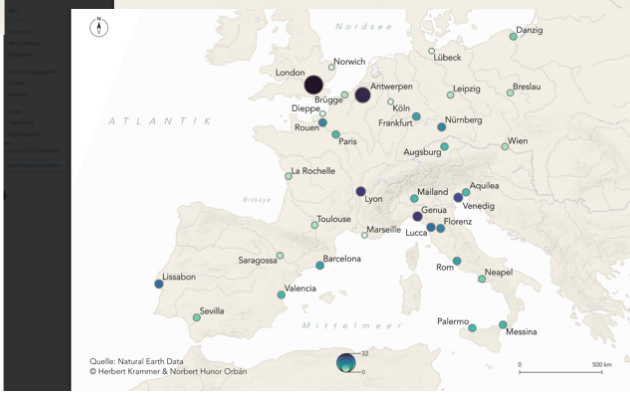
With the digital edition taking shape, the project is moving into deeper analysis:

- **Text Mining:** Using tools like **Voyant-Tools** and **AntConc** to perform similarity analysis on the texts, hoping to reveal clusters of transmission and patterns in terminology.
- **Network Analysis:** Visualizing the data to trace **trade routes** and commercial networks by mapping the cities mentioned in the reckoning examples.


Lucca
Pisa
Florenz
Konstantinopel
Rom
Palermo
Messina
Ceuta
Algier
Bugia
Alexandria
Mittelmeer
Schwarzes Meer
Biskaya


Leipzig
Köln
Altenburg
Eger
Frankfurt
Zwickau
Wien
Venedig
Konstantinopel
Algier
Kairo
Mittelmeer
Schwarzes Meer
Biskaya
Quelle: Natural Earth Data
© Herbert Krammer & Norbert Hunor Orbán


Köln
Eger
Prag
Regensburg
Nürnberg
Innsbruck
Wien
Ofen
Venedig
Konstantinopel
Rom
Algier
Kairo
Mittelmeer
Schwarzes Meer
Quelle: Natural Earth Data
© Herbert Krammer & Norbert Hunor Orbán


Nordsee
Danzig
Lübeck
Norwich
London
Antwerpen
Leipzig
Breslau
Brügge
Köln
Dieppe
Frankfurt
Nürnberg
Rouen
Paris
Augsburg
Wien
ATLANTIK
La Rochelle
Lyon
Mailand
Aquilea
Venedig
Genua
Florenz
Toulouse
Marseille
Lucca
Saragossa
Barcelona
Rom
Neapel
Lissabon
Valencia
Sevilla
Palermo
Messina
Mittelmeer
Biskaya
Quelle: Natural Earth Data
© Herbert Krammer & Norbert Hunor Orbán