



## Electronic Text Reuse Acquisition Project

GÖTTINGEN CENTRE FOR  
DIGITAL HUMANITIES

*Dr. Marco Büchler, Greta Franzini, Emily Franzini, Maria Moritz*

CLARIN-D Fach-AG Workshop, Leipzig, 30.06. – 01.07.2015



# Overview

## How

were paraphrases, allusions or translations **reused** by **authors** over **time**?

## Why

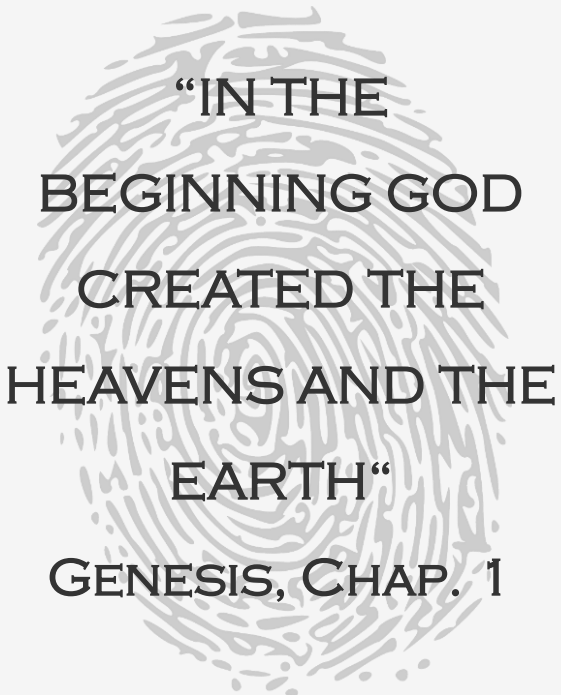
was **part** of the text reused? What influenced the text?

## Aim

Develop a **methodology** to „measure“ historical text reuse **in spite** of its **diversity**

## Through

Big Data, texts in Ancient Greek, German, English, Italian, Latin



**Information overload** (mass digitization) **vs. information poverty** (many lost works).  
**Incomplete cut-out** of what actually **exists & existed**

## Visualization Example

Bible graph visualization (TRAViz) to show variants of verses.

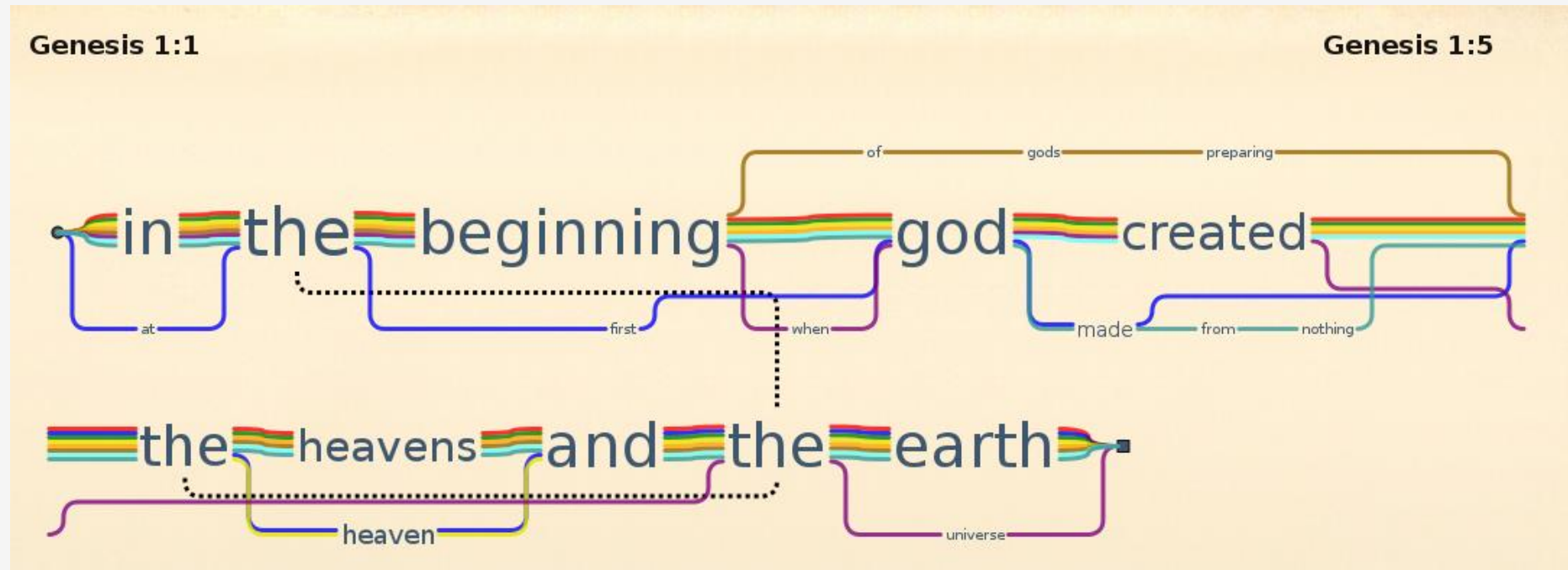
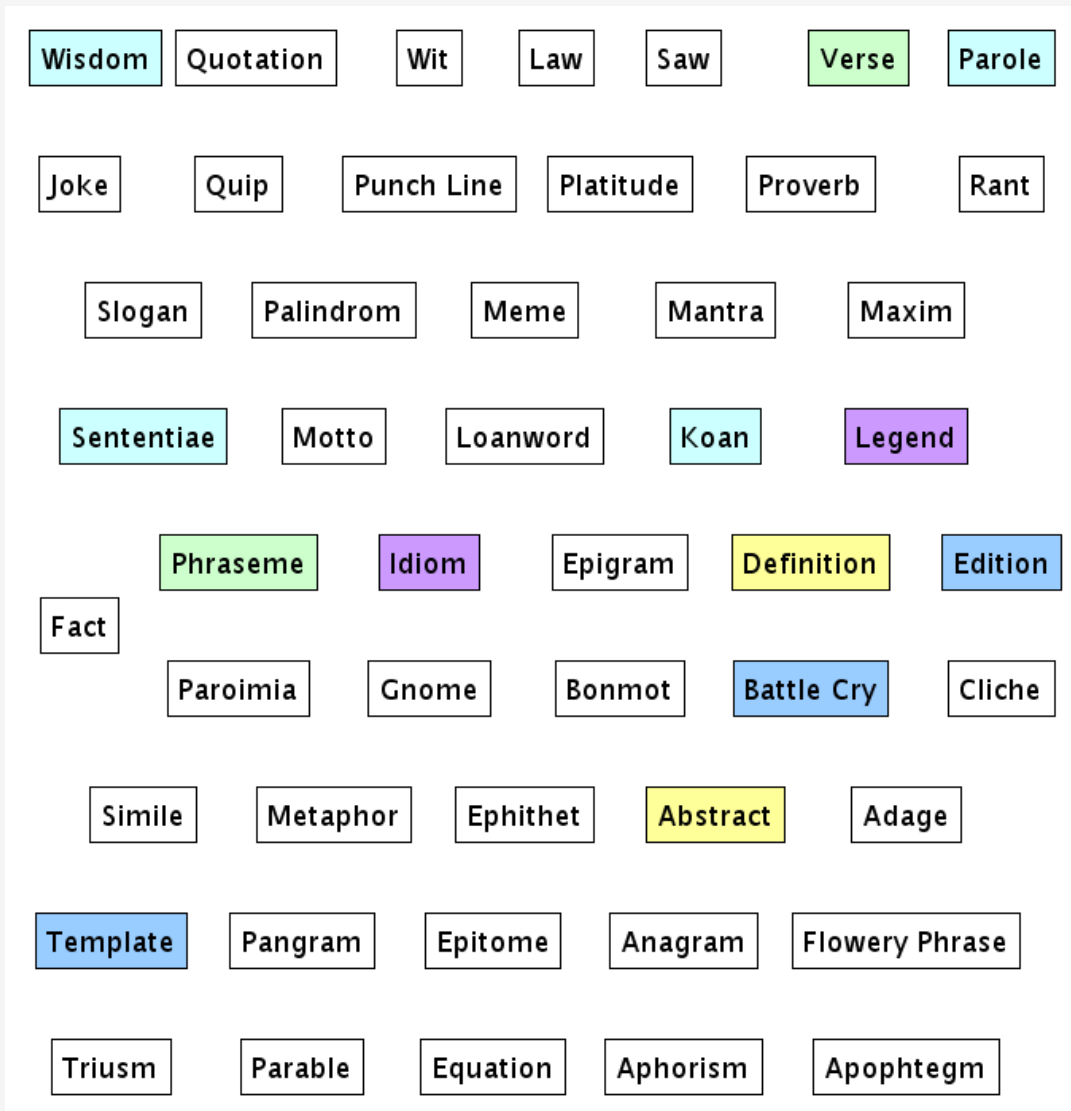


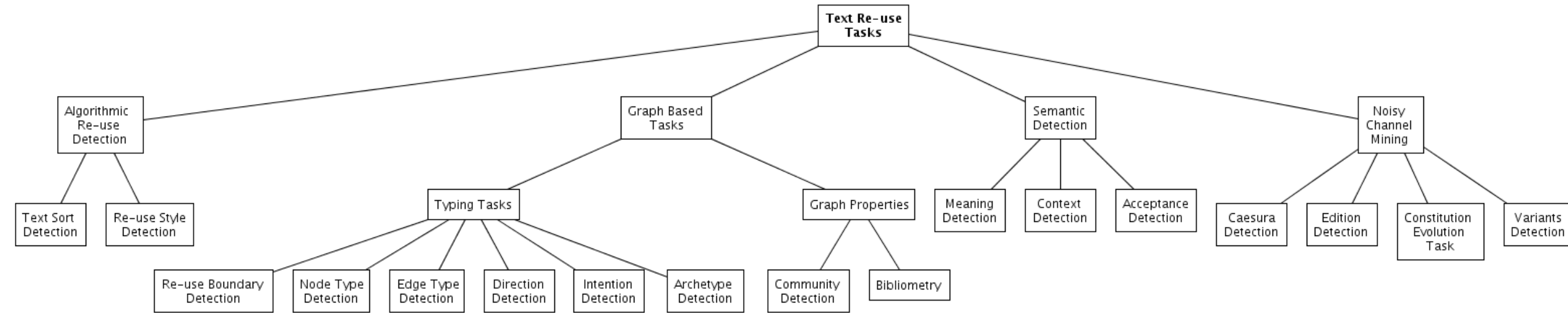
Image by  
Stefan Jänicke



- **Stability (yellow):** syntactic vs. semantic
- **Purpose/Intention (green)**
- **Size of Text Reuse (blue)**
- **Literary classification (light blue)**
- **Degree of distribution (purple)**

## Challenge

Language Modeling to cope with this diversity



Direction detection (not just symmetric relation)  
Boundary detection (not just point to the sentence)

## Challenge

Realistic time management. Be aware of the complexity of a task.

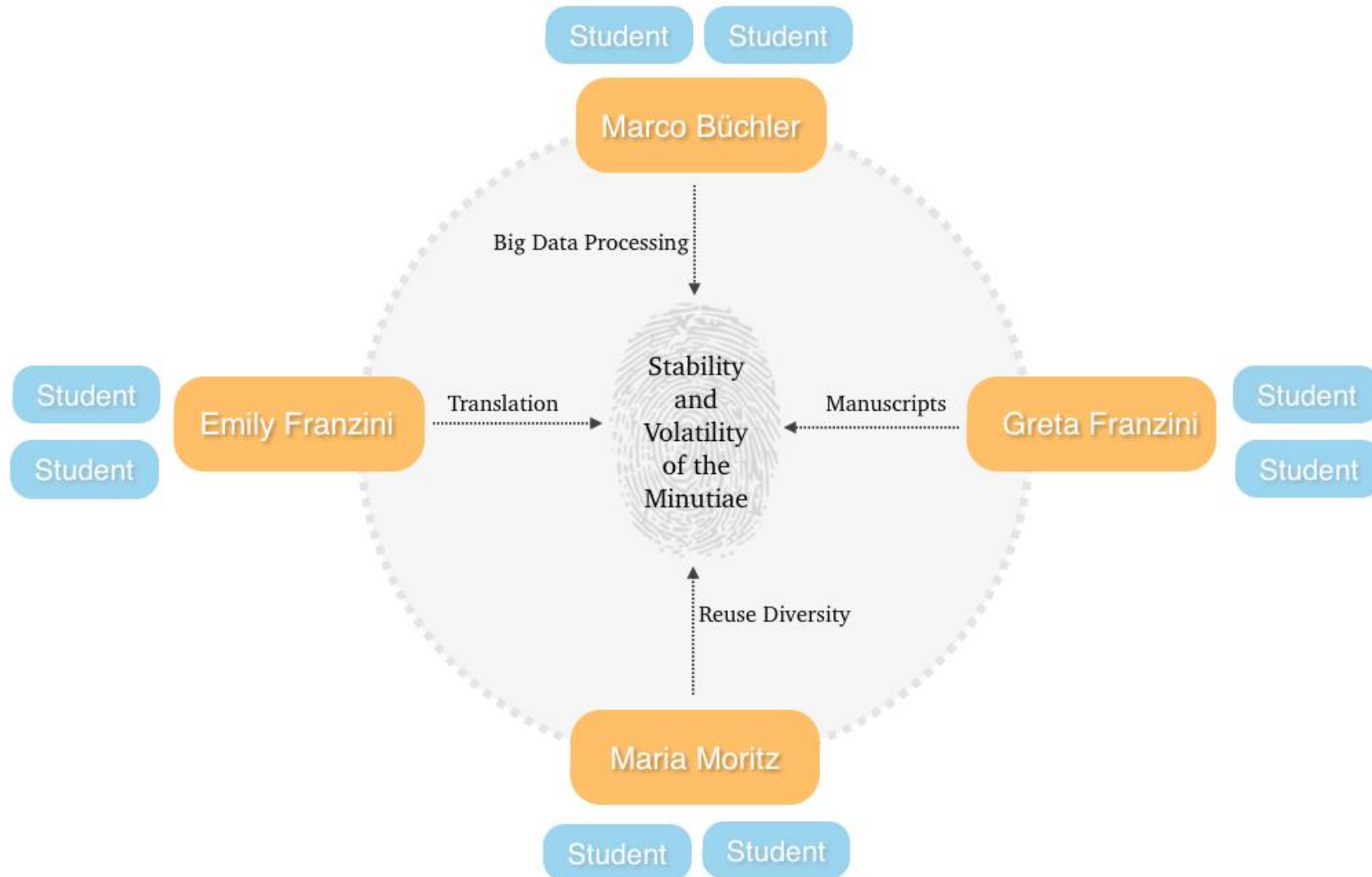
Different **formats**: annotation; encoding.

## Challenge


How to **homogenize** datasets in order to ensure interoperability and **minimize** data preparation?

cit-quote-bibl	blockquote	bibl without quote
<pre> &lt;cit&gt;   &lt;quote&gt;     du/o ku/nas a)rgoi\     ei(/ponto   &lt;/quote&gt;   &lt;bibl n="Hom. Od. 2.11"&gt;     Od. 2.11   &lt;/bibl&gt; &lt;/cit&gt; </pre>	<pre> &lt;quote rend="blockquote"&gt;   &lt;line&gt;     a)gxou= d' i(stame/nh e)/pea     ptero/enta proshu/da     &lt;bibl n="Hom. Il. 4.92"&gt;Il. 4.92&lt;/bibl&gt;   &lt;/line&gt;&lt;line&gt;     a)ll' a)/ge nu=n ma/stiga kai\     h(ni/a sigalo/enta     &lt;bibl n="Hom. Il. 5.226"&gt;Il. 5.226&lt;/bibl&gt;   &lt;/line&gt; &lt;/quote&gt; </pre>	<pre> &lt;p&gt;   [...]a)nti\ tou= proe/pinon. kuri/ws   ga/r e)sti tou=to propi/nein, to\   e(te/rw  pro\ e(autou= dou=nai   piei=n. kai ( *)odusseu\s de\ para\   tw=  *(omh/rw    &lt;bibl n="Hom. Od. 13.57"&gt;Od.     13.57&lt;/bibl&gt;   [...] &lt;/p&gt; </pre>

# Team

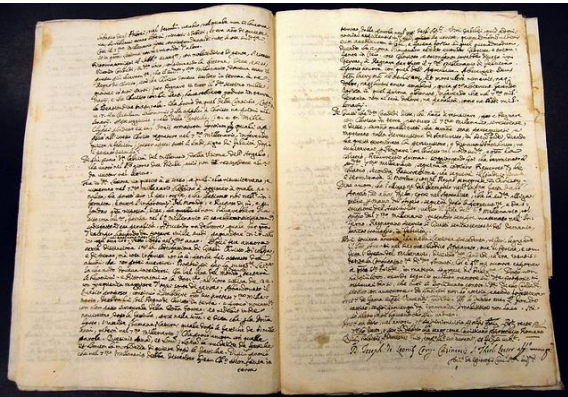


???

- 
- Which are the determinable **minutiae** of text reuse and which of them can be **calculated** by **machines**?
  - **What can't** be calculated?
  - How can text reuse analysis be improved by the calculation of **paradigmatic relations**?



# Greta Franzini, Humanist



???

Can the study of text reuse...



- ▶ ...tell us something about the **accuracy** and **confidence** of **variation**?  
Which authors quote the source more literally and why?
- ▶ ... tell us something about the **nature** of **transmission contamination**?

# Maria Moritz, Comp. Scientist

???



When observing a sentence or word that has been reused, one may ask:

- 
- 
- What are the **differences** and **similarities**, **what** do **they tell** us about the texts?
  - **How** does one **classify** the similarities and the differences (**Wittgenstein**)?
  - How and can those findings help to improve **NLP**?

# Emily Franzini, Humanist



???

The reuse of text **across languages** is a gold mine for **transcultural** studies.  
When observing **translated** text,...

- ...what is the scale of **divergence** from one text to the other?
- ...what is the **divergence caused by**? Translator competence or linguistic, cultural, ideological and political norms?
- Could **machine translators** be used to **rid** a **translation** of any **contextual influences**?



**Electronic Text Reuse Acquisition Project**

GÖTTINGEN CENTRE FOR  
DIGITAL HUMANITIES

*Thank you for your attention!*

*[etrap.gcdh.de](http://etrap.gcdh.de)*

