Model the source first! Towards Computer-Assisted Semantic Text Modelling and source criticism 2.0

David Zbíral • Robert L. J. Shaw • Tomáš Hampejs • Adam Mertel

Since the very beginning of the Dissident Networks Project (DISSINET) in 2019, we have been working on a **data model and workflow for the collection of complex structured data from textual sources**: principally, in our case, the records of inquisition trials. By doing so, we aim to deploy **various quantitative approaches** to the exploration and explanation of the **social, spatial, and discursive patterns** of **medieval religious dissent and inquisition**.

Our solution to this need is founded on one key observation: that, in historical research, we often deal with sources that are fundamentally **relational** in what they convey: that is to say, sources which contain extensive information concerning relationships and interactions among and between people, places, physical objects, events, meanings, etc. These connections can be transformed quite naturally into structured data.

On this basis, we developed an approach that we call **Computer-Assisted Semantic Text Modelling (CASTEMO)**. CASTEMO falls within the broader family of statement-based approaches to data collection, i.e. to modelling texts through statements with a syntactic structure whose basis is a relation between a predicate, a subject, and objects. The most comparable approach that we came across during our work on CASTEMO is Roberto Franzosi's Quantitative Narrative Analysis (see his *Quantitative narrative analysis*, Thousand Oaks: SAGE, 2010), an approach whose potential has been hugely underestimated in the social sciences and—to the best of our knowledge—completely neglected in history. With CASTEMO, we set out to make statement-based data collection both **more comprehensive** in terms of syntactic and semantic modelling and **more user-friendly** in terms of infrastructure.

CASTEMO is well-suited to any research where the precise rendering of the source's wording and context of any piece of information matters. It certainly can be used for selective data collection—i.e. capturing portions or elements of a source—but bears full fruit if used for a comparatively **maximalistic approach**. It is therefore often more time-intensive than simply extracting the data most obviously pertinent to a particular research problem, but it has immense rewards. Above all, it allows researchers to preserve a very high level of nuance and supporting detail: not just the "positive information", but also the **language of our sources** (specific terminology in original languages), **conflicting evidence**, information given in a

non-indicative modality, and the **conditions of production** of information (the order and context in which it is given, the interplay of questions and answers, etc.).

Systematically representing such information in the collected data not only allows us to do justice to the intricate interweaving of voices in historical writings, transcribed biographic interviews in social scientific research, and other types of texts, but also to pave the way towards a new practice of a **computer-assisted source criticism**.

Why structured data?

The main rationale for the use of structured data in historical and social scientific research is the discovery of **patterns** which could easily escape even very close unassisted reading. This is mostly because our normal practices of reading tend to prefer the richest anecdotal narratives, and patterns we already suspect to find or ones we want to question. **Computer-assistance in reading**, however, can provide **generalization, visualization, and summarization**, allowing the researcher to "zoom" in and out and discover significant patterns that are otherwise unanticipated.

For instance, our normal reading practices often uncover apparent **differences** between patterns of religious engagement within two dissident groups, or patterns of investigation of two different inquisitors, etc. But it is quite hard to assess **more than two or three sets of phenomena simultaneously**. Similarly, we cannot accurately measure the **proportion** of these differences. The collection and analysis of structured data offers precisely such possibilities.

To illustrate this, let us look at the broad picture of the importance of men vs. women in Kent heresy trials in 1511–12 and in the trial against the Guglielmites in Milan in 1300. In the parallel coordinates plot in Fig. 1, each line represents a person involved in dissident activities; the red ones are women, the blue ones are men. The line shows the scores of each individual on some of the standard measures of importance of an individual within a social network.

We clearly see that women do not seem so very important in Kent while they appear more important than men among the Guglielmites. That much would also be obvious to any attentive reader comparing the two sources without quantification and visual assistance. However, can we, with just pen and pencil, **work in parallel on three, four, seven outlooks on importance** in a community? That's already hard to do. And certainly, without structured data it is virtually impossible to find out things about **whole distributions** and **compare these distributions across datasets**. In this case, for example, it is now possible to see that in the Kent dataset there is a wide variety across the whole range of the centrality measures used; while in the Guglielmites dataset, there seem to be three outliers and then a core of individuals with broadly similar scores. To answer questions concerning distributions and proportions (including the proportions of causal factors), we absolutely need structured data and appropriate analytical methods.

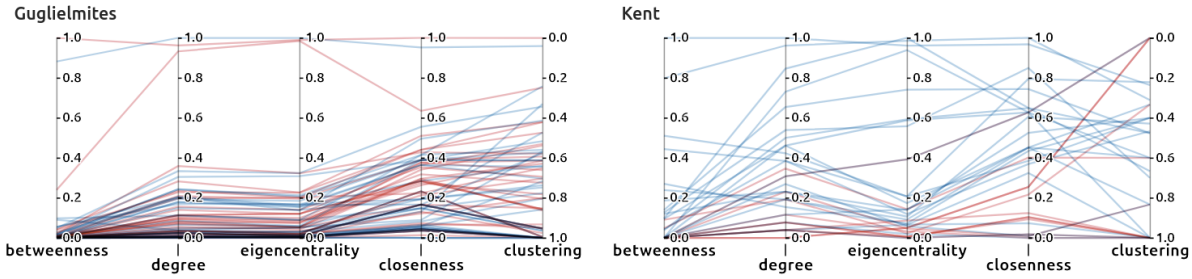


Figure 1. Parallel coordinates plot comparing the centrality scores of men and women across two datasets.

This is why we should consider extending our practices of reading towards what could be called **serial and scalable reading**.

Serial and scalable reading

In our proposal for "serial and scalable reading", the concept of seriality alludes to Pierre Chaunu's concept of **"serial history"**. Chaunu's vision revolved around using homogenous series of data for macrohistorical studies. Our take on seriality, however, is not limited by the boundaries of macrohistory. Rather, it is a matter of **treating large series of data within a coherent framework (data model)** in order to uncover broader patterns in our sources at every level of scale. We nevertheless remain within Chaunu's characterization of serial history as *"less interested in individual facts than in elements that can be integrated into a homogeneous series"* (Pierre Chaunu, "L'histoire sérielle: Bilan et perspectives", *Revue historique* 243 (1970), p. 297).

By **"scalable reading"**, we mean a reading which **transcends the dichotomy between close and distant reading**, neither of which describe well our procedure in DISSINET. Our reading and semantic encoding is definitely "close reading" in the sense of detail; in fact, we could even speak about "(very) slow reading", since coding sources our way as described further on easily takes 3-4 hours per standard page. At the same time, we are "close-reading" our sources in such a way exactly to achieve some of the aims of Franco Moretti's "distant reading", while still preserving the quality control of a historian. Overall, however, we want to achieve **"reading at all distances"**, providing new perspectives both for the "parachutists" and the "truffle hunters" (to borrow Emmanuel Le Roy Ladurie's proverbial image), as well as everyone in between. The ideal balance for today's historians is indeed to read across scales, to see the paths that lead from **small-scale relational patterns** all the way up to **big questions**. With scalable reading, we seek to transcend the notions of "close" and "distant", "macro" and "micro" by **zooming in and out**.

From exceptional anecdote to serial complexity

A new kind of serial history is especially relevant in the study of inquisition records. In this field, scholars have often made the equation between rich exceptional narrative and reliability. In fact, however, there is nothing in the formulaicity of our sources which would make them automatically unreliable, and detail and richness does not yet necessarily equate to truthfulness.

Indeed, **sources that appear formulaic in language and topic can actually contain unexpected complexity**. If superficially repetitive, the way that small scale features combine and relate to one another may be far less regular and hide unanticipated depths. Therefore, we strive to bring the study of inquisition records **beyond exceptional anecdotes** and instead focus on their **serial complexity**. Such an approach offers otherwise inaccessible insights and is by no means devoid of the nuance and beauty that historians appreciate.

And where has source criticism gone?

Nevertheless, in historical study, we usually deal with **very complex sources** full of vagaries, uncertainties, fuzziness, conflicting testimonies and so forth. Do computational approaches not run the risk of blinding us to such complexities?

In our experience, computational/digital history seems **strongly biased towards fact-oriented data collection**: that is to say, isolating key details from their original context in order to render them as data. And of course, such fact-oriented data collection has problems. Firstly, many important decisions on what represents a 'fact' have to be often taken *ad hoc* at the moment of entering a particular datum. Secondly, the data usually come without significant indication of the **conditions of their production**. This is not to say that researchers are unaware of the issues this creates: these conditions are often cited, usually in opening or concluding remarks. Nevertheless, they will never be adequately represented in the results themselves.

Take a sentence fairly typical of inquisition records which could read: *"Asked by the inquisitor, Peter said that he had seen Bernarda adore heretics in Lanta some twenty years ago."* For the purposes of computational study, it is tempting to transform this directly into "Bernarda adored heretics in Lanta in ca. 1225". But there is a substantial loss of information here. Who is speaking? To whom and in what context? What is the time span between the testimony and the reported event? All this is lost if we simply construct Bernarda's adoration of heretics around 1225 as a **fact** rather than a **claim within a source**.

In DISSINET, we think this is a major problem. We deal with sources which offer precious glimpses at the conditions of their production: sometimes we have a trace of the questions as

well as the answers within an interrogation, sometimes of the use of coercive procedures (e.g. detention or, more rarely, torture) used to help extract information, and sometimes of conflicting testimonies of the same events. In fact these conditions are often much more visible in trial records than in most medieval sources, and should provide the essential context for our analyses.

Since these conditions (e.g. interactions at trial) can also be neatly expressed as relations, our solution is to model them in the same way as the rest of our data, and indeed as part of the same data. Therefore **we at first model the source itself** rather than only isolating the details of immediate interest, literally **transforming it line by line into structured data**. This then allows us to inscribe source criticism at the heart of the analysis.

CASTEMO: a syntactic approach

Having made the case for **modelling sources in their entirety as structured, relational data**, let us now turn to *how* we seek to achieve this. Our process of source modelling produces data which are extremely close to the original but at the same time come **enhanced and formalized**.

Our approach is strongly **syntactic**: it is based on the **manual formalization** of the sentences of our sources into **statements** with **subject(s), predicate(s) and two objects**. We have preferred this basic structure—a **quadruple** —because even a very simple and omnipresent type of sentence such as "Peter received a gift from Elisabeth" requires four "slots": subject, predicate, and two objects rather than one. Such basic statements are very often extended through various modifiers (representing adjectives, adverbs, etc.), which we call **properties**. Those properties also come in the quadruple structure, and allow the same flexibility and complexity as standard statements.

These statements, with their syntactic structure and rich, explicit semantic relationships, are woven together into a knowledge representation known as knowledge graph. Our knowledge graph is close to the notion of "Linked Data", but its structure comes even closer to representing natural language than some standards, e.g. those which attempt to reduce sentences into semantic triples.

DISSINET data model in a nutshell

While it is not possible to present every detail of the DISSINET data model in just a few paragraphs, its basics can be broken down as below.

Statements and other Entities: the pieces of the semantic jigsaw

As already stated, CASTEMO is founded on **statements**.

- In our data model, a **Statement** is an **Entity** type, with a unique identifier.
- All Statements are tied to:
    - a **Territory**, i.e. a container for a CASTEMO representation of a text or set of texts, usually organized in a "folder" structure of individual parts with metadata assigned
    - a **Resource**, i.e. a specific source for that text (such as a specific OCR-ed edition, a manuscript or a transcript of an interview).

Statements have the purpose of relating other **Entities** to one another. The other **Entity types** in our data model besides Statements, Territories, and Resources are:

- **Action type**
- **Concept**
- **Person**
- **Group**
- **(Physical) object**
- **Location**
- **Event**
- **Value**

Each of these Entities also has a **unique identifier (URI)** in the database which holds the collected data (in our case, RethinkDB, a JSON document database).

When representing text, we usually render Entities in the **original source language.**

- If our text says that somebody "adored" (*adoravit*) the heretics, we do not, upon data collection, transform this into any modern interpretive concept, but rather keep the original Latin verb. Any interpretation occurs at a different epistemic level of the data, clearly marked-off from the original wording.
- Our entity lists thus contain entries in Latin, Middle English, Occitan, etc., with modern English serving as an analytical metalanguage.
- This is especially important for **Action Type** entities, which define the predicate of Statements, and **Concepts**, which serve to characterise many other entities (see below).

This allows **Statements** to relate entities in a way that **not only matches the meaning of the original sentence, but closely mirrors its language.**

Modelling complex textual semantics

At its most basic level, a **Statement** is built around the quadruple structure mentioned above: a predicate (an instance of an **Action Type**), which connects any number of entities in up to three **actant positions**—subject, first object and second object. The number of actant positions depends on the predicate's valency. Some predicates have no actants (e.g., "it rains"), some only have the subject (e.g., "John came"), some have subject and one object (e.g., "John brought some books"), and still others have subject and two object positions, each of which with different semantics (e.g., "John brought some books to Margaret").

To take a real example in the register of sentences of Peter Seila (Languedoc, 1241–2):

Roms de Sapiac … consuluit quosdam Valdenses pro infirmitate filii sui (in modern English translation: "Roms de Sapiac consulted certain Waldensians on the infirmity of his son").

At its simplest level:

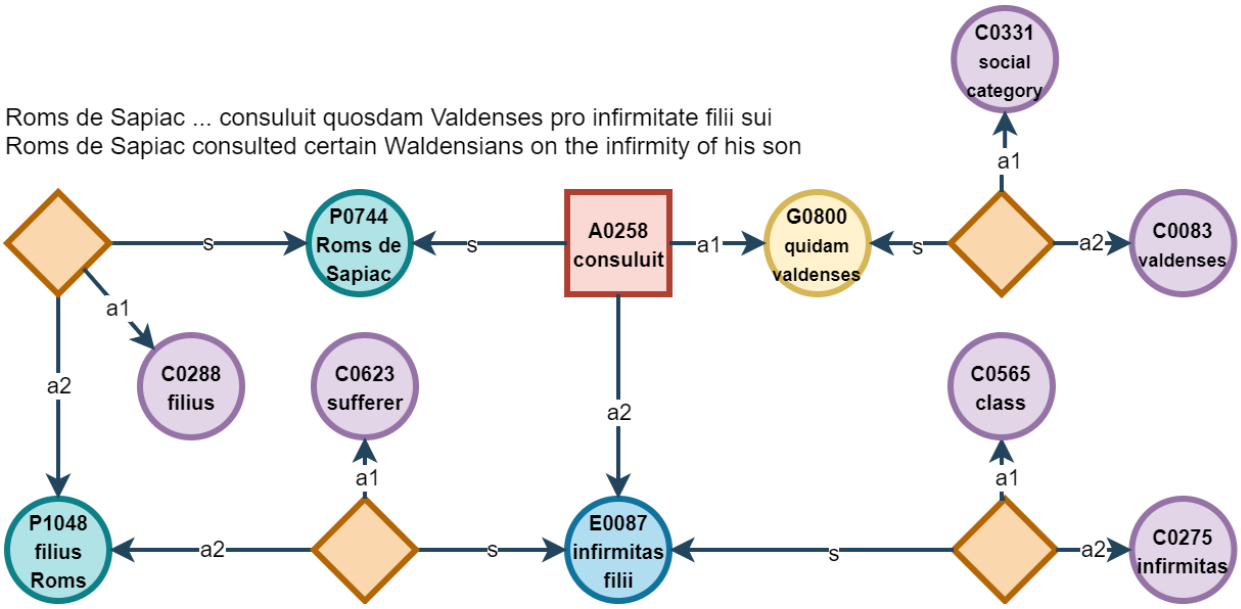
- "Roms de Sapiac" is the subject (a **Person**);
- the predicate is "*consuluit*", i.e. "consulted (an **Action Type**);
- object 1 is "*quidam Valdenses*", i.e. "some Waldensians" (a **Group**); and
- object 2 is "*infirmitas filii sui*", i.e. the "infirmity of his son" (an **Event**).

But there is much more complexity to model in this example.

- For instance, how to make clear that the **Event** labelled "infirmitas filii sui" relates to another person, Roms' son?
- How do we record that the specific group of "Valdenses" is an **instance** of the general concept of "Valdenses", so that information is inscribed reliably and unequivocally instead of being just a part of the label?

This is where **properties** come into play: these serve to expand the mother Statement it in various ways.

- Properties apply a **"property type"** (the kind of property to be attached) and a **"property value"** (the content of that property) to their subject.
- The subject of the property can be any of the entities involved in the Statement – its actants (typically with adjectives), or the action itself (typically with adverbs and adverbial clauses).



Properties and their uses

The uses of properties include:

- **Instantiating** Entities to corresponding Concepts.
    - For example, once you create a specific "apple" as an Object, you also want to record that it is an instance of a wider class of things (in this case, the generic Concept of apple).
    - These Concepts themselves are arranged in conceptual hierarchies inside our data model (e.g. "apple" is also a subset of "fruit", and "fruit" of "food", thus allowing you to find any mention of food of whatever type in the database).

- Modelling **additional descriptors** (e.g. **adjectives** and **appositions)** concerning actants.
    - These are modelled through Concepts (for property types and qualitative property values) and Values (for instance, for numerical property values).
    - Crucially, **the properties remain attached to the textual context from which they derive**, rather than the descriptors simply being recorded as facts about the Entities in question.

- Modelling **relationships expressed without a verb (e.g. via** apposition or by something's very definition) concerning actants.
    - For instance "son *of Roms*" is linked to "Roms" via a property attached to the latter: he possesses a "son" known as "son of Roms".

- Defining **time** and **place of action** as far as possible.
    - This can be in absolute or relative terms (e.g. in relation to another Statement).
    - It can utilize fuzzy descriptors as well as more defined values.

- Recording **other adverbials,** for example those concerning **manner of action, circumstances, causes or consequences of action**.
    - It is possible to relate the actions of Statements to adverbial Concepts and Values, in the same manner as adjectives and appositions.
    - It is also possible to relate the actions of Statements to other Statements through properties. This is often essential in matters of causation: in coding a sentence beginning "Because of this deed…", it will be necessary to create a link back to the "deed" described in a preceding statement.

Statement perspectives: Modality, Epistemic level, and Certainty

To further characterize the claim denoted by a Statement, we carefully record three different aspects of **perspective**: modality, epistemic level, and certainty.

- **Modality** is intrinsic to the text and describes in what semantic mode the statement is formulated. This allows us, for instance, to differentiate **positive from negative** assertions ("Peter was there" vs "Peter was not there"), **assertions** from **questions** ("Was Peter there?"), **wishes** ("May Peter go there"), and **conditions** ("Were Peter to go there…"), and so on.
- **Epistemic level** describes the position from which a Statement is formulated. We differentiate three levels: **textual** (that is, an explicit claim of the source), **interpretive** (that is, our interpretation but still close to the text), and **inferential** (that is, our inference external to the source). Importantly, we are able to mark epistemic levels for both whole statements and their individual parts. (This comes in handy for example if the text provides a property value, e.g. "baker", but the property type, e.g. "profession", is an editorial classification.)
- **Certainty** is, in our understanding, the editorial judgement on how the statement in question is reliable.

Modelling textual order and information flow

Statements can be like a jigsaw puzzle in and of themselves. But, as Entities, they also form part of a bigger picture, informative for the way in which information was created and communicated.

Crucially, our data model preserves the relation of any statement not only to the entire **Territory (text or set of texts)**, but also its specific **part** (e.g., a specific document within a

specific trial which is in turn a part of an inquisitorial register). This hierarchical structure, with folders and subfolders going all the way down to any level you like, serves to model the embeddedness of documents. We always provide Territory parts with **metadata** concerning the roles of Persons in relation to them. Thus it is immediately clear in whose deposition in front of which inquisitor a specific piece of information appears. Within this structure, the exact textual order of the information is also preserved by the order of the Statements.

Our line-by-line coding, meanwhile, records the full cascade of information flow.

- Our sources contain many semantics that concern the flow of information. Questions ("After being asked whether…"), responses ("He responded…"), admissions, claims, and denunciations ("She said…", "She confessed…", "He accused…").
- These are critical to situating the origin of the information offered by our sources: they must be recorded in their own time, place, and context.

Our structure allows for **chains** to be formed between multiple statements, in order to capture the flow of information fully and precisely. E.g., when coding a passage such as "But next day he heard the aforesaid Raymond Peter telling him that that night the aforesaid heretics had hereticated the said sick man [Raynaldetus of Soricino]" (*Sed in crastinum audivit [Petrus Pictavini] prædictum Raymundum Petri referentem sibi quod illa nocte prædicti hæretici hæreticaverant dictum infirmum [Raynaldetum de Soricino]*), we not only want to record the illicit action, but also the subsequent hearsay and finally the deposition (Fig. 2).

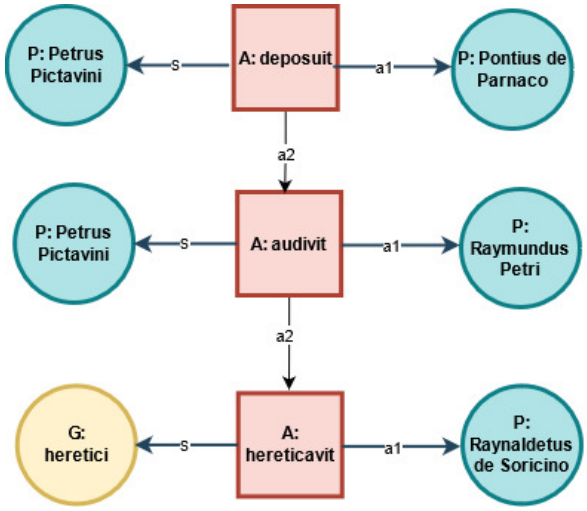


Figure 2. Modelling all the reported stages of information flow.

Carefully nesting any piece of information in the appropriate lowest-level Territory part and preserving the information flow gives our data a layering that is still extremely rare in historical datasets. Both these elements help us to complete our goal of producing data that is highly contextualized within the discursive qualities of the evidence. The pitfalls of

"fact-oriented" data collection are almost completely avoided, and crucially, such source modelling helps us not only to achieve an extremely high level of source-critical nuance to the quantitative study of medieval dissidence, but to make the information flows, including the process of trial and recording, an object of study in itself.

Interface

The tour of our data model makes clear our **maximalistic approach** to data collection and its possibilities. We preserve the utmost detail of the original, and at the same time we enhance it with a lot of additional information. It allows us maximum flexibility in terms of **data projections**, **exploratory data analysis**, and asking otherwise unapproachable questions. We can dig out much more than if our data collection was very hypothesis-driven, and therefore very selective.

It is also, as previously stated, inevitably quite time consuming and intensive work. Interface development has thus run parallel to our efforts to hone our data model, and indeed informed it.

In developing an interface, we started simple. We used a set of interconnected **Google Sheets tables**: for statements, "coding sheets" where each row represents one statement and the columns serve to place the different entities in the correct "boxes" (Fig. 3); for each type of entity such as persons, locations etc., another dedicated table.

It turned out that starting simple and building up was actually almost the only possible way of starting at all. Using this sort of 'nuts and bolts' interface to attempt source coding allowed us to refine not only the requirements for a future front-end, but, most crucially, the needs of the data model itself: The data model as it stands is testament to this approach of trying to structure ever more perfectly real source data in an ad-hoc interface of columns: if we had started from a pre-existing tool or standard for data collection, inevitably tied to somebody else's ontology, it would not have attained its present flexibility. **Constant feedback between our emergent coding practices and the challenges we discovered in the sources**—and crucially the **constant team discussion of these issues**—has shaped our entire approach to source modelling.

id	parent_id	epistemol	text	id_subject	subject	id_action_or_r	action_or_relation	action_or_relation_english	id_actant1	actant1	id_actant2
T107-1-02-010			Interrogatus	P0216	Andreas Saramita	A0126	"interrogatus/a" [judicialiter]	interrogated on trial (by sb - [whether / concerning st])	P0092	Guido de Cochenato	T107-1-02-011
T107-1-02-011	T107-1-02-010		si aliquis de parentela sua tam ex parte patris quam ex parte matris	G0009	parentela Andree Saramite	A0192	"fuit"	was (sb/st)	C0006 #C0019	hereticus/a [generaliter] #credens hereticorum [sc. Catharorum]	
T107-1-02-012			respondit	P0216	Andreas Saramita	A0080	"respondit"	replied (to sb - st)	P0092	Guido de Cochenato	T107-1-02-013
T107-1-02-013	T107-1-02-012	2-in	non	G0009	parentela Andree Saramite	A0192	"fuit"	was (sb/st)	C0006 #C0019	hereticus/a [generaliter] #credens hereticorum [sc. Catharorum]	
T107-1-02-014			Interrogatus	P0216	Andreas Saramita	A0126	"interrogatus/a" [judicialiter]	interrogated on trial (by sb - [whether / concerning st])	P0092	Guido de Cochenato	T107-1-02-015
T107-1-02-015	T107-1-02-014		si cognovit Guillelmam, sepultam apud monasterium Clarevalis,	P0216	Andreas Saramita	A0162	"novit"	knew (sb/st)	P0218	Guillelma (in vita)	
T107-1-02-015.4		2-in	sepultam	P0218	Guillelma (in vita)	A0108	"decessit"	died		#VALUE!	
T107-1-02-015.6			sepultam apud monasterium Clarevalis	O0005	corpus Guillelme	A0187	NULL	has location	L0056	monasterium Clarevalis	
T107-1-02-016			respondit	P0216	Andreas Saramita	A0080	"respondit"	replied (to sb - st)	P0092	Guido de Cochenato	T107-1-02-017
T107-1-02-017	T107-1-02-016	2-in	quod sic	P0216	Andreas Saramita	A0162	"novit"	knew (sb/st)	P0218	Guillelma (in vita)	
T107-1-02-017.4		3-in	et invenit quod ita erat	P0216	Andreas Saramita	A0032	"credidit"	believed / had faith (in st)	T107-1-02-017.6	T107-1-02-017.6	
T107-1-02-017.6	T107-1-02-017.4		quod fuit filia quondam regis Boemie, ut dicebatur	P0218	Guillelma (in vita)	A0106	"filia"	daughter (of sb)	P0222	(rex Boemie)	

Figure 3. DISSINET data collection in Google Sheets.

We have now transitioned to a **more coder-friendly web-based data collection interface** of our own creation: [InkVisitor] (Fig. 4). This allows us to perform exactly this kind of rich semantic coding, but in a much more convenient way than Google Sheets. InkVisitor is a very general data collection interface, with a wide range of potential applications beyond the coding of inquisition records. Both the interface and the underlying data model are appropriate to other source categories, whether they be historical in origin or modern, e.g. interview transcripts gathered by researchers in the social sciences. In the latter field, the use of InkVisitor and the underlying CASTEMO data model to capture information represent an emerging alternative to Computer-Assisted Qualitative Data Analysis Software.

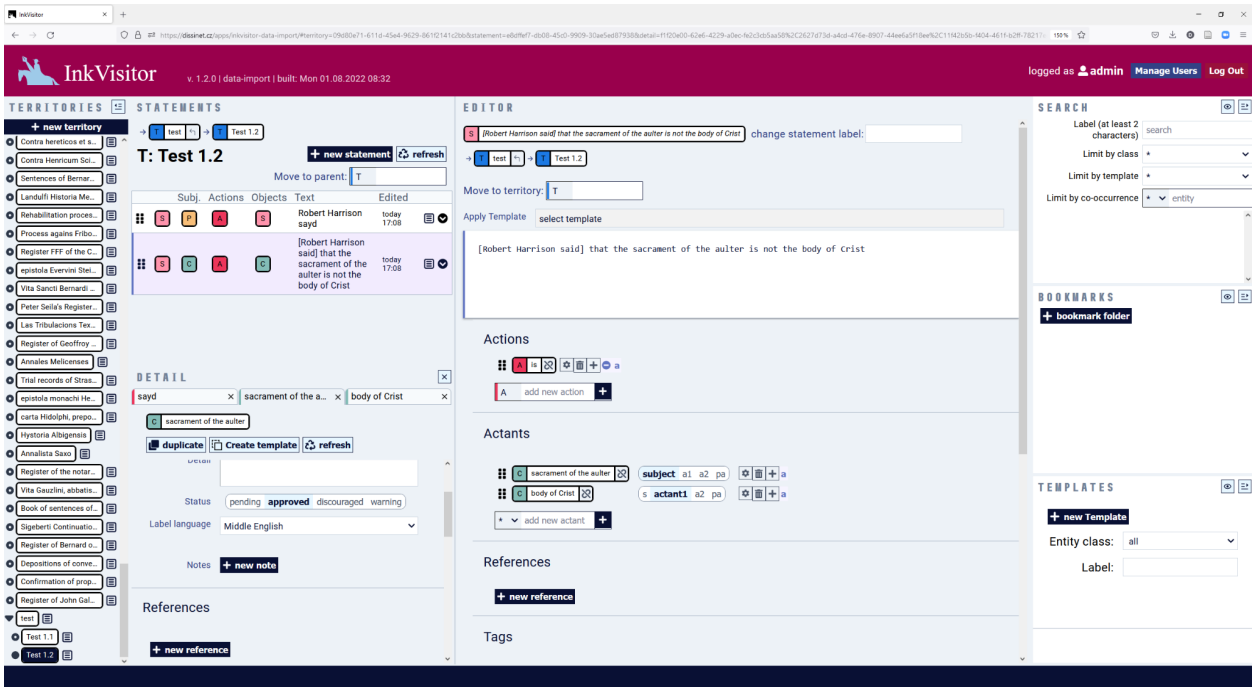


Figure 4. InkVisitor interface for data collection.

Wrapping up

This article can only skim the surface of Computer-Assisted Semantic Text Modelling and its rationales. We hope, nevertheless, that the potential of CASTEMO stands clear. It produces data which model, in the first instance, the syntactic-semantic structure of the texts themselves, in order to then address all kinds of research questions where perspective, context, and discourse matter. Our modelling of sources enables **"serial and scalable reading"**: by seeing many data points together, we are able to appreciate unanticipated complexity in both complex narratives and more formulaic texts, and to allow **bigger patterns to emerge from small-scale, local relations**. We can not only re-approach textual information in new ways, but also directly address the texts themselves. If computational history sometimes appears less critical in its approach to sources, our approach to data collection allows for what might be called "**source criticism 2.0**". The classic source critical focus on the conditions of information production and transmission can now be bolstered by systematic data on those conditions and by analytical techniques which allow us to face the challenges of our sources of information as never before.

The crucial pay-off of source modelling is that it allows us to represent the sources in their entirety, and make them themselves the objects of systematic computational analysis. Instead of falling prey to first impressions or rich but anecdotic narratives, we are able to defer some

decisions on the specific data projection to the moment when we have already closely inspected the source at various scales. We thus make rigorously **informed decisions** regarding our analyses. They are informed not just by closely reading the source in a classical sense, but by understanding it at a new level, one which is virtually inaccessible to unassisted reading practices.

Acknowledgements:

The research presented in this article is a part of the "Dissident Networks Project" (DISSINET, https://dissinet.cz) and received funding from the Czech Science Foundation (project No. GX19-26975X "Dissident Religious Cultures in Medieval Europe from the Perspective of Social Network Analysis and Geographic Information Systems"). The article has been revised with the aid of funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101000442).

To cite this article:

David Zbíral; Robert L. J. Shaw; Tomáš Hampejs; Adam Mertel (2022). Model the source first! Towards Computer-Assisted Semantic Text Modelling and source criticism 2.0. *Zenodo*. DOI: 10.5281/zenodo.6963579.