The STructured Assertion Record (STAR) Model for Event-based Representation of Historical Information

Background
Among the many challenges posed by the study of the world of the Christian communities of the eleventh century is the scattered and arbitrary nature of the surviving sources; following from this, the information we derive from our sources varies wildly in its precision and its reliability. Moreover, our sources represent a wide variety of viewpoints that are represented therein and give rise to an equally wide variety of scholarly interpretations, which can be very difficult to represent usefully in collections of digital data.

One of the goals of the RELEVEN project has been to re-think the way in which we model data about the past. A straightforward Linked Open Data (LOD)-based model tends to result in positivistic data collections, where the pieces of information become separated from their provenance and context and thus difficult or impossible to verify (Kahn and Simon 2020); on the other hand, models such as factoid prosopography (Bradley and Pasin 2015) or even Wikidata (‘Wikidata’ 2023) emphasize the retention of provenance (source) information, at the expense of providing coherent and non-contradictory accounts of past events (Baillie 2021).

Methodology
One of the key problems we face for eleventh-century history, which informs our approach to a solution, is that the historical claims that are made are often not traceable in their entirety to specific primary source material. A suitable information system must represent claims that arise from a preponderance of evidence, claims that arise from absence of evidence, and even claims that arise from conjecture.

This constraint leads the the core of our solution, which is to extend the linked open data model from a triple (subject, predicate, object) to a five-pointed structure (thus “STAR model”) which includes subject, predicate, object, source, and authority. This model allows for the possibility of adopting and enriching existing LOD collections with source and authority data when it exists, since the subject, predicate, and object can come from any defined ontology.
A schema of our model is presented in Figure 1. Its backbone is the CIDOC-CRM (Bekiari et al. 2021), using the ‘E13 Attribute Assignment’ entity as the core of each assertion. The subject and object are then classes of E1 CRM Entity and the predicate is reified as an E55 Type; we then model the assertion as having been P17 motivated by the source (if a specific source exists) and P14 carried out by some actor, who functions as the authority. Where the source is a textual document, we draw on the LRMoo (formerly FRBRoo) ontology (Bekiari et al. 2021) to create a more in-depth description of the source passage in question. The use of LRMoo also allows the aggregation of assertions into statements; this is a useful feature to map between common narrative statements such as “The emperor had the rebel executed by drowning” and the collection of assertions that arise from the multiple CIDOC-CRM triples needed to represent such a statement.

On top of this model, we provide a small custom ontology to represent the perspectives of different actors where these perspectives can be modelled (these are almost invariably the authors of primary or secondary historical works, i.e. the authorities for the various assertions in the database. A perspective is made up of one or more statements; for each statement, the perspective can either accept it or reject it. This modelling of rejection in particular provides a solution for the common problem of how to represent information derived from an absence of evidence, or from statements that can only be expressed as negatives. Moreover, the event-based nature of CIDOC-CRM allows for these perspectives to have a temporal aspect; that is, we can represent the perspective of an actor as it develops or shifts during their lifetime.
With the STAR model we have been able to ingest data from the factoid-based Prosopography of the Byzantine World, and have begun a similar exercise to ingest data from the Pleiades gazetteer. Future work will include data-mining experiments for the automated detection of conflicts in information, and methods for visualization of the conflicts and ambiguities that are detected.

Acknowledgments
This work arises from a project that has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant agreement No. 101002357).

References