Representing Humanities Research Data using Complementary Provenance Models

Bridget Almas[2], Monica Berti[2], Sayeed Choudhury[3], David Dubin[1], Megan Senseney[1], Karen M. Wickett[1]

1 Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign
2 Perseus Digital Library, Tufts University and University of Leipzig
3 Data Conservancy, Johns Hopkins University

Overview

The Systematic Assertion Model (SAM) is a formal account of research data and its content - relating data identity, change over time, and varieties of scientific or scholarly equivalence directly to research transactions. Humanities use cases include a breadth of computational and interpersonal activities that call for a more inclusive understanding of provenance than SAM alone can provide. Integrating complementary views of the same research enterprise by combining SAM entities and properties with those from other models such as W3C's PROV and Open Annotation (OA) provides a richer, more contextualized view of data encoding and use in humanities research.
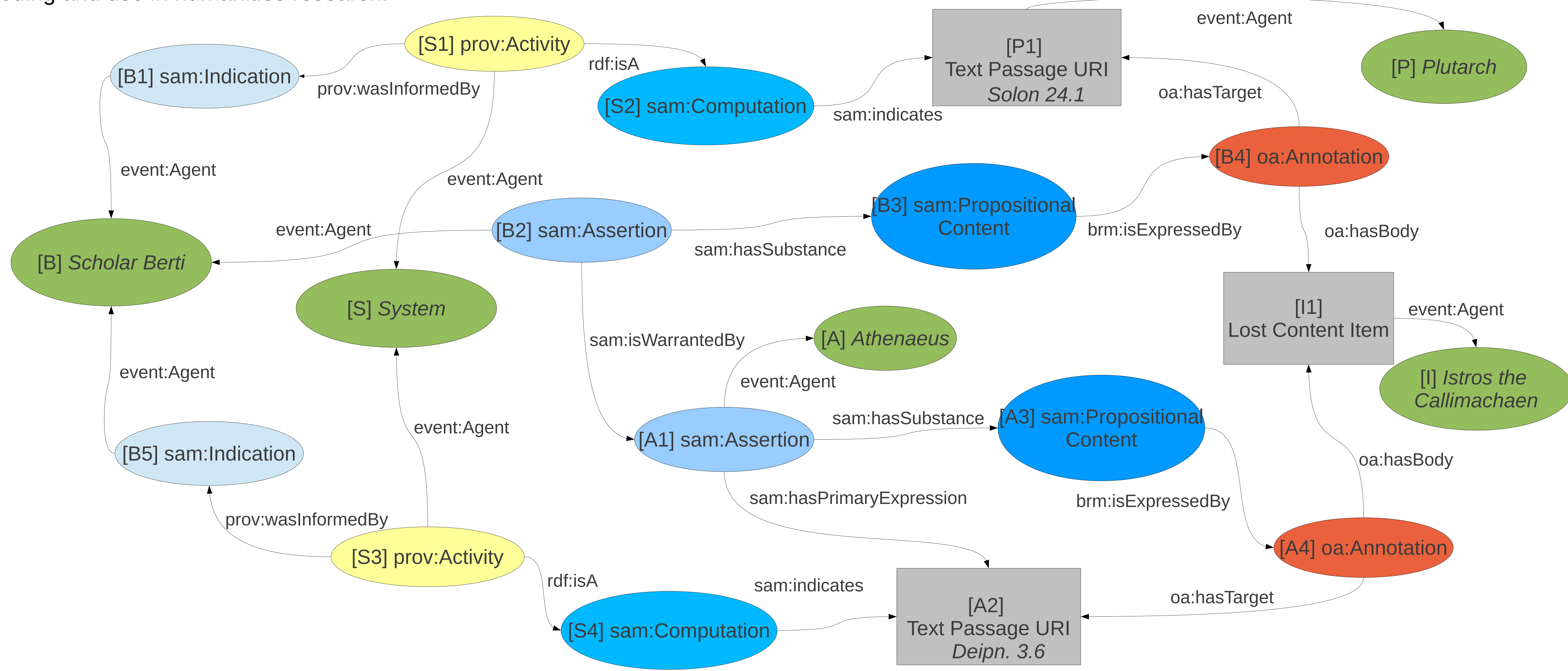
Use Case Details

Scholar Berti [B] is the *agent of* an Indication [B1] – the selection of a string of text. [B1] *informs* an Activity [S1] by the system - a Computation [S2] of the URI for a text passage [P1] by Plutarch [P].

[B] makes an Assertion [B2] the *substance of* which is Propositional Content [B3] as *expressed by* an Annotation [B4] *targeting* the Plutarch text [P1] as a reuse of a Lost Content Item [I1] attributed to Istros the Callimachaen [I].

[B] *indicates* another string of text [B5] that *informs* an Activity [S3] by the system - a Computation [S4] of the URI for a text passage [A2] by Athenaeus [A]. The text at [A2] is the *primary expression of* an Assertion [A1], the *substance of* which is Propositional Content [A3] as *expressed by* an Annotation [A4] *targeting* [A2] as a reuse of Lost Content Item [I1] attributed to Istros.

B's Assertion [B2] *is warranted by* A's ancient Assertion [A1].

SAM Indication

Event in which some abstract structures (e.g. a series of characters) are indicated in the pursuit of some intellectual or creative goal.

SAM Assertion

Event in which an agent advances a claim. The primary expressive may be a natural language sentence or any abstract arrangement of symbols. Assertions that are warranted by observations or computations are *systematic assertions.*

SAM Propositional Content
Language-independent bearer of truth values

SAM Computation

An event that may reflect contingencies such as scholarly interpretation. Does not necessarily imply *creation*. E.g. an analytic process may indicate a string of text which already existed.

OA Annotation

Expresses the relationship between two or more resources, including metadata about the relationship concerning creation and intent.

PROV Activity

Event through which entities come into existence and/or change to become new entities. Activities are dynamic aspects of the world, such as actions, processes, etc.

Summary: A scholar wants to annotate what she thinks is a quotation or text re-use of a lost text attributed to Istros in an extant source text by Plutarch (a reference to the origin of the word 'sycophants' where Istros is not named). To substantiate her argument, the scholar must also identify corroborating material, including instances of Istros' text in other primary sources, in this example that of Athenaeus Deipn. 3.6 (who does name Istros as the source).

Conclusion

Combining complementary data provenance models enables us to more precisely track and document shared resources, ultimately improving data quality and encouraging further sharing. Using PROV Activities we share precise details about system actions and processes leading to the extraction of text and creation of URI endpoints. We use SAM to identify the contingent aspect of the underlying resources as things which are subject to interpretation and which were in existence prior to their use as data in our analysis. We use OA to share concrete serializations of the analyses in the form of annotations.

Libraries: Transformation of Humanities (LTH) is an IMLS-funded project focused on building a network of collections of resources for the humanities and creating tools to facilitate contributions to research from a wide range of users.

INSTITUTE *of* Museum *and* Library SERVICES
LG-06-11-032611

DataConservancy
At The Sheridan Libraries, Johns Hopkins University

Perseus Digital Library
Tufts University Medford, MA, USA
The University of Leipzig Leipzig, Germany

CIRSS

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois