Archaeological research questions as a basis for semantic data modeling and querying

Analyzing 388 questions from conference abstracts to understand what archaeologists really ask and what IT systems must support

Pavlos Fafalios[a,b], Athina Kritsotaki[a], Christianna-Despina Pratikaki[a], Martin Doerr[a]

[a] Centre for Cultural Informatics, FORTH – Institute of Computer Science (Heraklion, Greece)
[b] School of Production Engineering and Management, Technical University of Crete, University Campus (Chania, Greece)

Contact: fafalios@ics.forth.gr

The Challenge: Bridging Archaeology and IT

- **Research begins with questions!**
  - Archaeologists ask about relationships, change and meaning
  - A useful **research infrastructure** must **connect** heterogeneous evidence, not merely store isolated finds
- Digital research infrastructures now enable global data aggregation and integrated access for researchers worldwide through **shared ontologies**
- But a critical question remains unanswered:

Can existing data models and information systems adequately accommodate the extreme diversity of archaeological data?

The Opportunity

Projects (like ARIADNE) already use ontologies (CIDOC CRM) to aggregate archaeological data at scale across institutions and borders.

The Gap

Model adequacy can only be proven through systematic analysis of real, concrete research questions, not assumptions!

The Design Requirements

- Research questions expose three design requirements:

Ontology coverage

Which entities and relationships must the data model represent?

Query access

Which recurring question patterns must be expressible and reusable?

Reasoning support

Where do statistics, correlation and interpretation enter the workflow?

The Problem: Real Questions are Scarce

- Concrete archaeological research questions are surprisingly difficult to collect
- Several structural barriers make systematic gathering nearly impossible:

→ Institutional Privacy

Museums register user queries but usually avoid disclosing them (competitive and legal sensitivity)

→ Cognitive Erasure

Scientists often forget specific questions immediately after receiving an answer (the question dissolves once solved)

→ Funding Language

Grant applications contain speculative, high-level questions (far too vague to implement in IT systems)

→ No Central Repository

To the best of our knowledge, no shared collection of concrete archaeological research questions currently exists anywhere.

Our Approach

- Reconstructing Questions from Research

The Source

121 abstracts from the 19th European Association of Archaeologists conference (Pilsen 2013), from the section "*Interpreting the Archaeological Record" (Sessions A01-A08)*.

This material is representative of the research trends of that year!

The Method

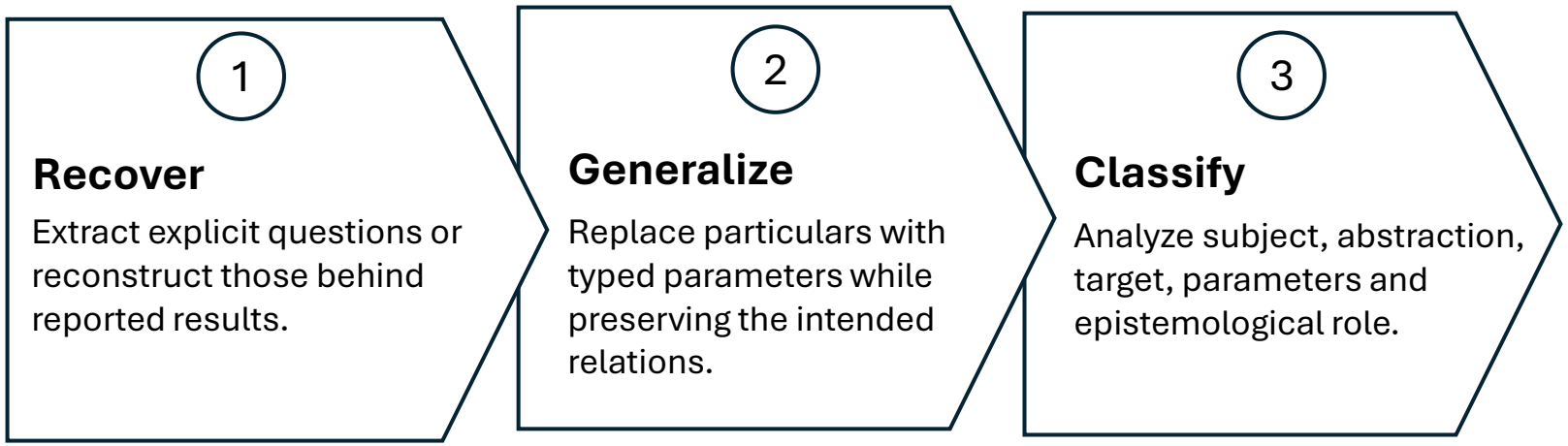
Research results and conclusions were systematically reverse-engineered into the specific, answerable questions that motivated them.

The Output

The reconstructed questions were transformed into parametric, generalizable question forms, suitable for evaluation against IT system capabilities.

Methodology

- Three-step method



Methodology

- From one question to a reusable query pattern

Specific question

What are the exotica which were found in the protohistoric settlement of Roca?

Parameterized form

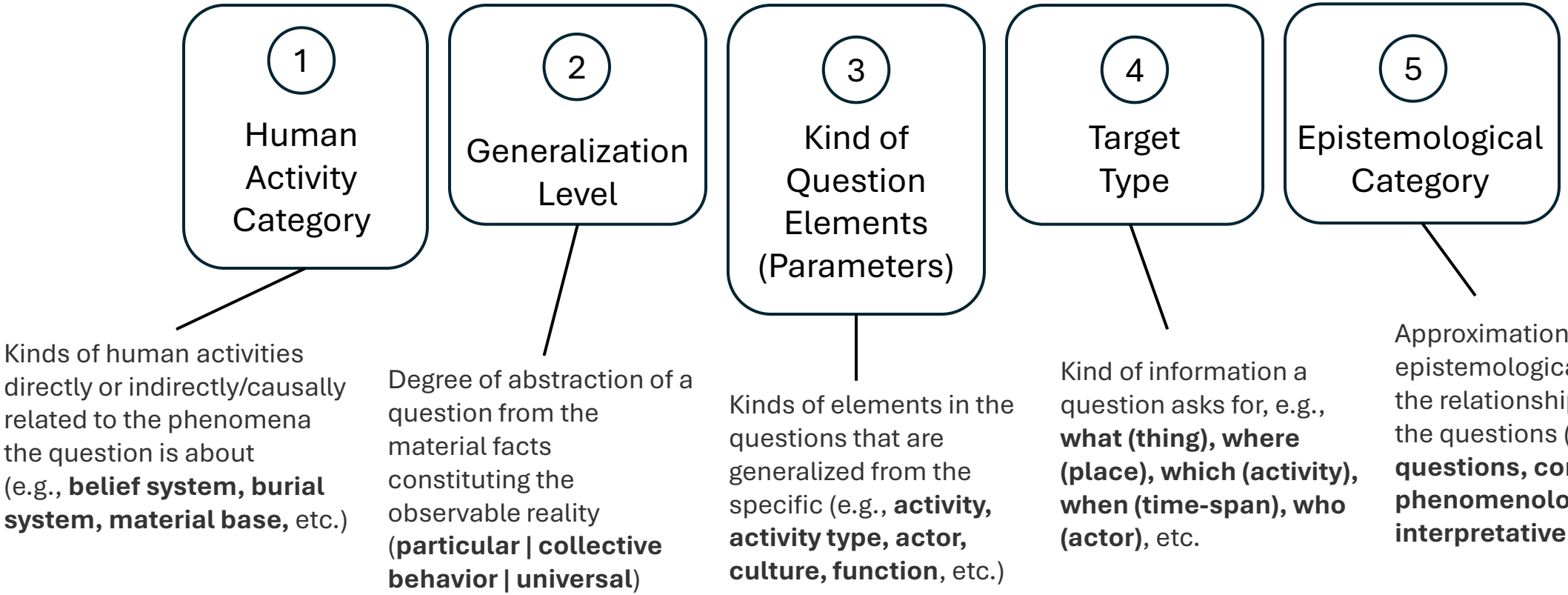
What **[object type]** found in the **[built context]** of **[period]** at **[place]**?

*Note: The relation **'found in'** still requires formal semantic mapping!*

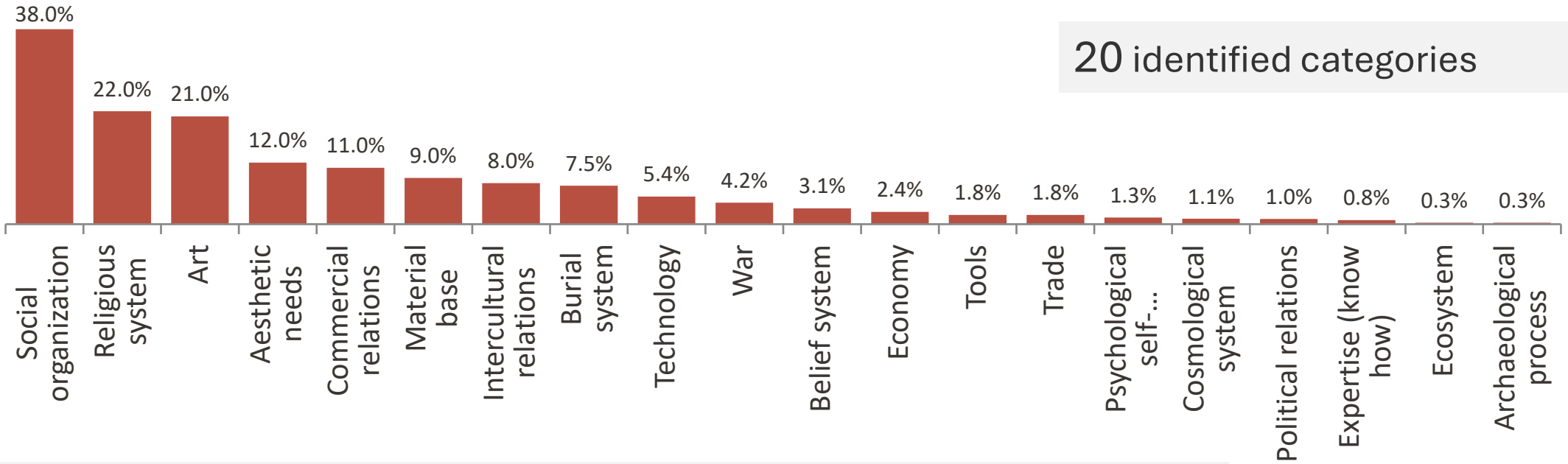
In total, 388 questions were extracted or reconstructed from the 121 abstracts!

Methodology

- Five classification dimensions



Findings – Human Activity Categories



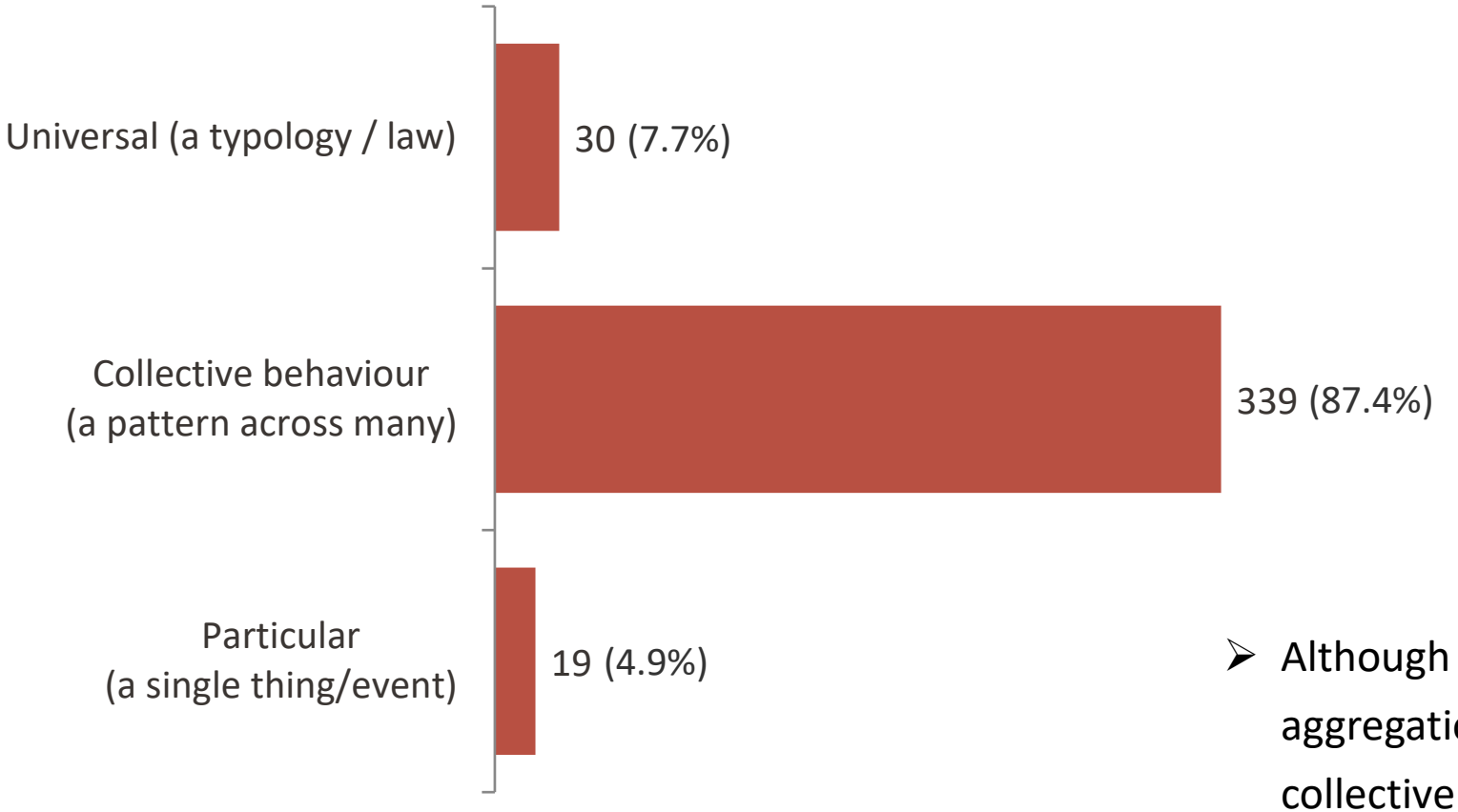
20 identified categories

Interrelated Categories that stand out:

- Social Organization (38%) and Intercultural Relations (8%)
- Religious System (22%), Burial System (7.5%), and Belief System (3.1%)
- Art (33%) and Aesthetic Needs (12%)
- Commercial Relations (15%), Economy (2.4%), and Trade (1.8%)
- Material Base (9%), Technology (5.4%), Tools (1.8%), and Expertise (0.8%)

Burial evidence is abundant in excavations, yet ranks much below social organization (7.5%)

Findings – Generalization Levels

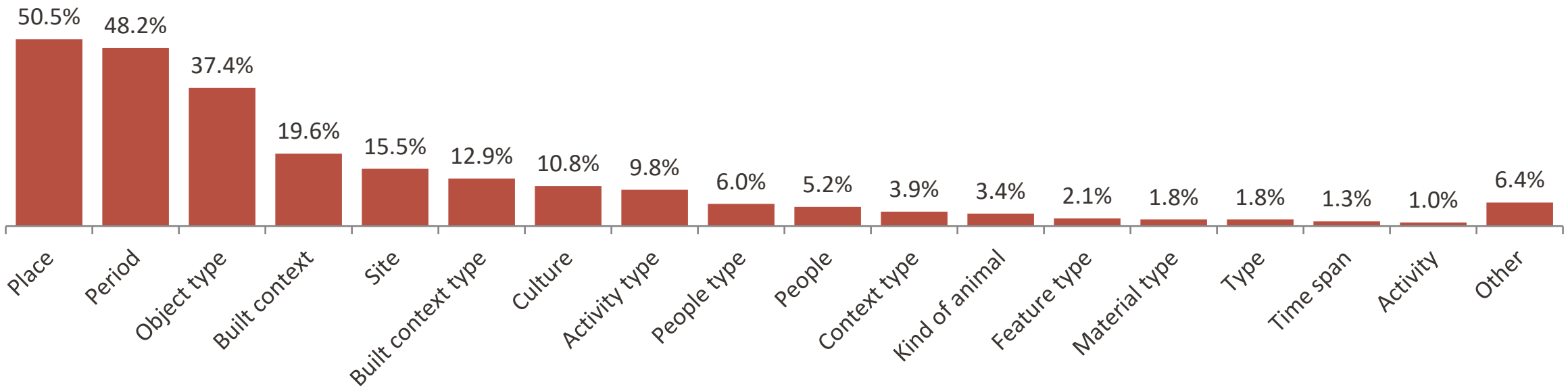


87.4%

of questions target collective behaviour (patterns inferred across many particulars), not a single find.

- Although query languages (SQL and SPARQL) support aggregation, expressing archaeologically meaningful collective patterns often requires domain-specific modelling, reasoning, and analytical layers!

Findings – Kinds of Question Elements (Parameters)



- **Place** alone constrains half of all questions
- **Period** and **Culture** together cover more than 50% of the questions (the classic *spatiotemporal* anchor)

- **Object type (**37%) is the dominant "type" parameter; **material type** is surprisingly rare (1.8%)
- Surprisingly, **explicit time-spans** are very rare!

Findings – Target Types

The classical *who, when, where, what* questions cover about half of the questions.

The "when" question is rarely asked directly ("which time-span" with 3.9%), but the questions about activities or events amount to 10.9% and may imply a "when" something has happened.

- **Overall, the findings demonstrate the importance of events and activities in the archaeological discourse**

Whereas questions for events or activities are easy to program, further research is needed to find out if or how IT could support answering **"how" or "why" questions**!

Findings – Epistemological Categories

31%	5.5%	41.5%	22%
Descriptive questions	**Comparison of specific types**	**Phenomenological correlations**	**Interpretative conclusions**
Listing facts directly from the source records	Similarity, difference or co-occurrence of types	Correlating categories of types, no stated cause	Inferring cause, actor, or other non-evident relations

- Phenomenological correlations dominate: archaeologists most often seek a pattern between two categories of things, without yet asserting a cause. **This is exactly the reasoning step current query languages support least!**
- Similarly, questions related to interpretive conclusions are difficult to answer by the usual SQL-like query systems!

Three Questions That Expose the Challenge

- Concrete examples from the corpus. Each requires more than simple fact retrieval

Evidence & contacts	What evidence gives us the information that there were contacts linking Central Europe with the highly developed civilization of Mycenaean culture? **[Abstract A01.08]**
Comparison of cultures	What distinguishes the Mycenaean culture from other East-Mediterranean centers, with respect to the kinds of objects? **[Abstract A01.12]**
Comparison & interpretation	Which is the difference between burial rituals and the social status of Iranian shamans to those of shamans of Siberian type? **[Abstract A03.16]**

What These Findings Mean for Semantic Modeling & Querying

1. **Model collective behavior, not just particulars**

   87% of questions concern collective behavior inferred across multiple instances. Supporting such questions requires more than retrieving individual facts: it requires domain-aware aggregation, statistical analysis, and interpretation.

2. **Prioritize Place, Period, and Object Type as core parameters**

   These are the three most frequent parameter; they should be first-class, well-indexed access paths.

3. **Support event- and process-centred queries**

   How/why and activity/event questions are common and central to archaeological discourse, yet record-oriented schemas (such as Dublin Core) provide limited support for representing rich event structures and their interrelationships.

4. **Represent the epistemological step, not only the subject**

   Phenomenological correlation and interpretative-conclusion questions need documented reasoning chains: intermediate steps that a system can support even if the final inference cannot be automated!

Conclusion: A reusable corpus, not a final answer!

https://isl.ics.forth.gr/archaeological_questions/

- 388 empirically grounded questions, openly published, for reuse by archaeologists and IT/DH researchers.
- **Findings reveal a systematic gap between current IT systems** (strong on particulars; who/what/where) **and how archaeologists actually reason** (collective behavior, correlations, causes).
- Our methodology is deliberately scoped to "evaluating the record" (post-excavation), not fieldwork; a different, complementary question space!

- Next step: use this material to **extend a formal ontology** (e.g., CIDOC CRM) and **implement IT systems with query mechanisms** that capture these query requirements!

Limitations

- Four limitations should guide the interpretation of the findings:

→ Limited coverage

Based on 121 abstracts from one EAA conference section; it provides a snapshot of that year, not archaeology as a whole.

→ Scope boundary

Focuses on evaluating the archaeological record (post-excavation); findings may not generalize to fieldwork questions.

→ Abstract-level evidence

Abstracts condense research aims and omit methodological detail; therefore, questions guiding the research may not appear in the published text

→ Requirements, not evaluation

The findings reveal modelling and querying requirements, but do not evaluate whether particular ontologies, query languages, and/or IT systems can answer the specific questions (**this is our next step!)**

Takeaways

- **Question-driven design** changes both the data model and the query layer
- Support **event-** and **process-centered** queries
- Model **collective behavior**, not just particulars
- Preserve the **reasoning chain** and the **documented evidence**
- **Extend** formal, **event-centric ontologies** (such as CIDOC CRM), and develop **IT systems** with the needed querying capabilities

Thank you!

Questions?

Explore the material

https://isl.ics.forth.gr/archaeological_questions/

Presenter:

Pavlos Fafalios (fafalios@ics.forth.gr)

FORTH INSTITUTE OF COMPUTER SCIENCE

centre for cultural informatics

TECHNICAL UNIVERSITY OF CRETE