

Challenges of Knowledge about the Past

…and semantic extraction from texts

Martin Doerr

Olympia 2026

Heraklion, Crete November 2026



Data about the Past

Data about the past *serves the reconstruction and understanding of* ***real "possible pasts".*** *It comes from* ***observation reports*** *and* ***extent material evidence*** *[1].*

We cannot go back: *Data about* ***world states*** *are always incomplete!* ***Quality*** *depends on completeness of* ***all available*** *data (in contrast to nomothetic sciences!). Strong* ***cross-disciplinary*** *relations, incl. analytical sciences. (impressive innovation rate)*

=> Need for global information integration, continuous accumulation, a sort of

Enterprise Knowledge

Typical disciplines: *History of all disciplines, archaeology, anthropology ("CH"), biodiversity, palaeontology, palaeogenetics,* ***epidemiology****,* ***ecology, geology****, forensics,…*

Data about the past is ***vital information*** *for the* ***current state*** *of the world in all scientific and scholarly aspects and for* ***extrapolations*** *into the near future, depending on human intervention.*



Reasoning

Reasoning *is very complex, comprehensive, using all possible methods, inductive-approximative. About* ***Contextual*** *knowledge = reconstructing facts, inferring impact factors, causes, interpolating evolutions, trends.* *(nomothetic sciences seek context* ***independent*** *knowledge)*

Basic method *is: "Inference to the Best Explanation" , (well-explained in scientific realism, since 2010s, e.g., by James Ladyman [2] and others).*

Each new fact ***reduces*** *possible pasts****,*** *improves or even overthrows explanations. => very high* ***knowledge revision*** *rate.*

Verified provenance *and material* ***evidence*** *is vital. Danger of circular argumentation, pseudo-independent evidence, guesswork, fraud.*

(e.g. chicken bones used for decades in geology for dating layers, and the latter by biologists for dating chicken bones, causing errors up to millions of years. Detected by ***reanalyzing museum specimen****)*



Our Respective Experience

Conceptual modelling *and KR-base implementations for (cross-domain) integration of reliable, precise data about "the past" and records of scientific observations..*

European ***research infrastructure*** *projects in Cultural Heritage, archaeology, biodiversity (also national), geology, KR-based integration in biomedicine, but also for information systems for* ***museums, archives****.*

Focus on ***empirical verification*** *of models by domain experts as ontological commitment, ("deep knowledge engineering"), in contrast to the still prevailing constructivist agenda.*

Leading the development and extensions of the ***CIDOC CRM****, ISO 21127:2006-2023, with the International Council of Museums* ***for 30 years.*** [3] *(translated to 7 languages. Besides others, national standard in P.R.China.)*

Conceptual basis *for most of our projects.*



Characteristics of CIDOC CRM

*Starting in **1996** as extensible "enterprise model" for **cultural-historical** data of museums and **natural** history, LAV, **query containment** via superproperties (key references work of D. Calvanese et al.)*

***Event / activity** centric, classes are **only for justifying** properties, rich **subproperty** hierarchies. Now nearly 1000 properties in extensions, **few top-level** concepts.*

***Domain-overarching** common concepts, relevant for global integration, ontology **harmonization** between domains (e.g. biodiversity-archaeology, OGC standards [4], libraries' FRBR-LMR)*

***Textual definitions**, about 200 – **1000 words** per class, 50 – 400 words per property. Enhanced, **revised** over years following **domain expert requests**. Complete evolution history. (FOL rules as far as the precision is empirically adequate)*

*Focus on **human action**, people, things and information **meeting in spacetime**.*

Meetings in Spacetime

(can be queried by 4 top-classes, 5 top-properties)



"LAÖKOÖN"
(copy)
(in Vatican museum)

Winckelmann's death

"…noble simplicity, silent grandeur…"
(in a library)

Winckelmann writes….



time

1755

unknown Roman

Winckelmann sees "Laökoön"

Johann Joachim Winckelmann
(archive information?)

"LAÖKOÖN"

unknown Roman copies "Laökoön"

Published Inference (in a library?)

Winckelmann's birth

Winckelmann's mother
(archive information?)

Greece Rome Germany *space*



Creating Knowledge Bases of Historical Data

Museum databases *including* ***natural history,*** *are mostly for collection management, or object type-specific (e.g., coins, "corpora"). Normalized to* ***few*** *fields.*

Collections *are separated by subject (e.g., porcelain),* ***contextual*** *relations between them often unknown. => In-house Enterprise Knowledge problem*

Curatorial contextual knowledge *is in notes, exhibition catalogues, publications [5]. Curators may contact some 500 museums for one exhibition, extensive research in* ***archives****. Global Enterprise Knowledge problem.*

Mapping *legacy bases to KR-bases is of* ***limited value*** *for interpreting historical facts.*

Manual data entry *from text to KR-bases does* ***not scale****. IT engineers still "hardwire" data entry forms. Generic KR entry tools are often cognitively or ergonomically* ***inadequate*** *or simply too complex.*



Semantic Extraction to CRM RDF

a first, few days experiment with Claude

*Digesting **all** CIDOC CRM SIG **community documents**: textual definitions, RDFS modules & examples, mailing list, issues discussed, minutes, publications.*

*Creating a 15-component **hybrid AI** / Python code, **verifying** AI arguments against **original** community texts, and **validating RDF** code.*

*Applied to **20 Wikipedia** pages, **supervised**. The surprise:*

*- created already better CRM instances **than IS students** after a CRM course.*

*- The **LLM can relate** the rich CRM **textual** definitions and community discussions to particulars in texts. (hardly possible with FOL rules only. Untrained experiments with CRM RDFS only were useless).*

*- The **event structure** and conceptual coverage of the CIDOC CRM itself appears to guide the extraction process to the **relevant** facts (but Claude may have applied additional relevance judgement).*



Semantic Extraction to CRM RDF cont'ed

*Applied to **20 Wikipedia** pages. The surprise:*

*- The Claude transformation comments were CRM knowledge **revision aware**: Claude about the tapestry of Bayeux:*

"**E84 Information Carrier** — the obvious class for "a physical thing that carries information" — turns out **not to exist** any more. It is absent from `crm_list('CRMbase')` entirely, and `crm_concept('E22')`'s debate list shows **why**: thread t1054, "Deletion of E84 Information Carrier and redistribution of examples," 2018-01-04, decided."

*It appears feasible to **stabilize** high quality interpretations of CRM concepts for semantic extractions in a **local model**, probably to be adapted for each NL, expert jargon etc.*

*Creation of **very large** high quality KR-bases about the past from **relevant textual sources** suddenly become possible!*



Challenges

OCR of "primary sources": *chronicles, archival records, traveller reports (Maghellan,…), exhibition catalogues, observation reports etc.*

*- **Relevance criteria** (see Doerr and Low 2010 [5] and "Significance 2.0" [6] , my favorite!).*

Supervision of AI interpretations:

*- Most tools for **reading** complex RDF junks are **cognitively inadequate**.*

*- **Workflows** of verification, global **coordination** of experts ("cataloguing rules",training)*

Scalable KR-bases:

*- Current platforms **not sufficient**. Competing with LLM contents? Distributed?*

*- Managing distributed **knowledge revisions**, **schema updates. Europeana** needed **7 years** for **one** schema update.*

***Adequate querying** for incomplete data, **Open World**, **deeply indirect** connections, explorative, knowledge **revision aware**. SPARQL is **cognitively inadequate**.*



Thank you!

Questions?

References:

[1] Martin Doerr, 2025. “Identifiable Individuals and Reality: Describing the Past by Formal Propositions”. ACM Journal on Computing and Cultural Heritage, Volume 18, Issue 2 Article No.: 27, Pages 1 – 32 https://doi.org/10.1145/3715011

[2] James Ladyman. 2002. Understanding Philosophy of Science. Routledge (August 6, 2012) e-book format ISBN 9780203463680

[3] CIDOC CRM Special Interest Group. 2021. Definition of the CIDOC Conceptual Reference Model version 7.1.1. Chryssoula Bekiari, George Bruseker, Martin Doerr, Christian-Emil Ore, Stephen Stead, Athanasios Velios (Editors). ICOM International Committee for Documentation (CIDOC) https://doi.org/10.26225/FDZH-X261

[4] Gerald Hiebel, Martin Doerr & Oeyvind Eide, Ø.2017. “CRMgeo: A spatiotemporal extension of CIDOC-CRM”. International Journal on Digital Libraries 18, 271–279 (2017). https://doi.org/10.1007/s00799-016-0192-4

[5] Martin Doerr and Tyan Juey Low, 2010. “A Postcard is not a Building”, ICOM Conference 2010, Shanhai, P.R. China. https://www.ics.forth.gr/_publications/CIDOC_2010_low_martin.pdf

[6] Roslyn Russell and Kylie Winkworth, “Significance 2.0 A guide to assessing the significance of collections”, Published: 28 May 2024, Collections Council of Australia Ltd, https://www.arts.gov.au/publications/significance-20