Ontology-Guided Synthetic Data Generation for Low-Resource Information Extraction: A Case Study in IT Heritage Domain

Nakanyseth Vuth[1*], Emrick Poncet[1*], Benjamin Lecouteux[1]
Caroline Djambian[2], Didier Schwab[1], Gilles Sérasset[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG | [2]Univ. Grenoble Alpes, GRESEC
38000 Grenoble, France
{first.last}@univ-grenoble-alpes.fr

tl;dr

We generate schema-consistent NER/RE training data directly from an ontology — no pre-existing knowledge base required — which we use to bootstrap cold-start IE in a niche domain (IT Heritage).

Motivation - the cold-start problem

- Expert annotation is costly in specialized domains (IT, Legal, Defense, etc.)
- 80 gold documents can't cover 18 entity types and 15 relations
- Distant supervision scales annotation but introduces significant label noise
- ***Reverse-IE**** avoids the noise, but needs an existing **knowledge base**
- In practice, specialized domains have neither **annotations** nor a **knowledge base** — only an ontology

Contribution - ontology as generative priors

- Ontology-first pipeline: no pre-existing knowledge base required, only a formal schema
- Full control over data generation: steer the sampling toward any underrepresented entity type or relation
- Applied to IT Heritage domain (18 entity types, 15 relations)

  → +19 NER F1, +15 RE F1 over gold-only

Methodology - the pipeline

Verbalization fidelity:
94.63% entities · 93.45% triples

Stage A: Extract working Schema

Ontology: 329 entity types, 86 relations

Working Schema: 18 entity types, 15 relations

Stage B: Sample a KG motif (per document)

Step 1: Randomly select an Anchor Entity Type (Protoganist)

Step 2: Sample degree and schema valid relation
k ~ Poisson(λ=2) → 3 outgoing edges

Step 3: Sample tail entities from candidate pool based on re-use prob α

Step 4: Inject Entity Instances

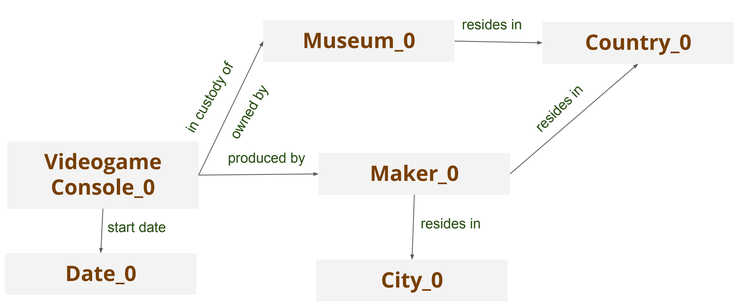
** Repeat until desired n triples achieved for a given KG

Stage C: Verbalization

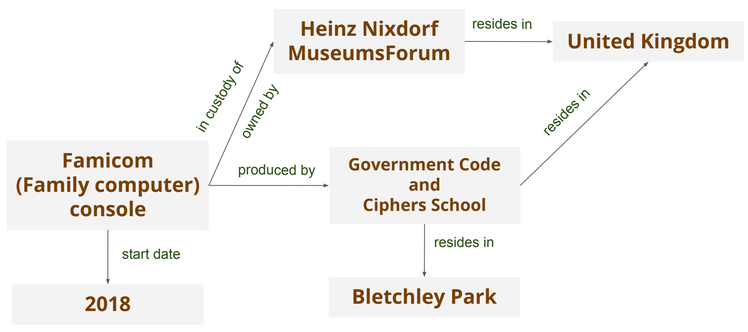
KGAST (Vuth et al. 2024)

Worked Example:

**The entity combinations in the Realized KGs are not necessarily factually accurate



KG Motif



Realized KG

Generated Text

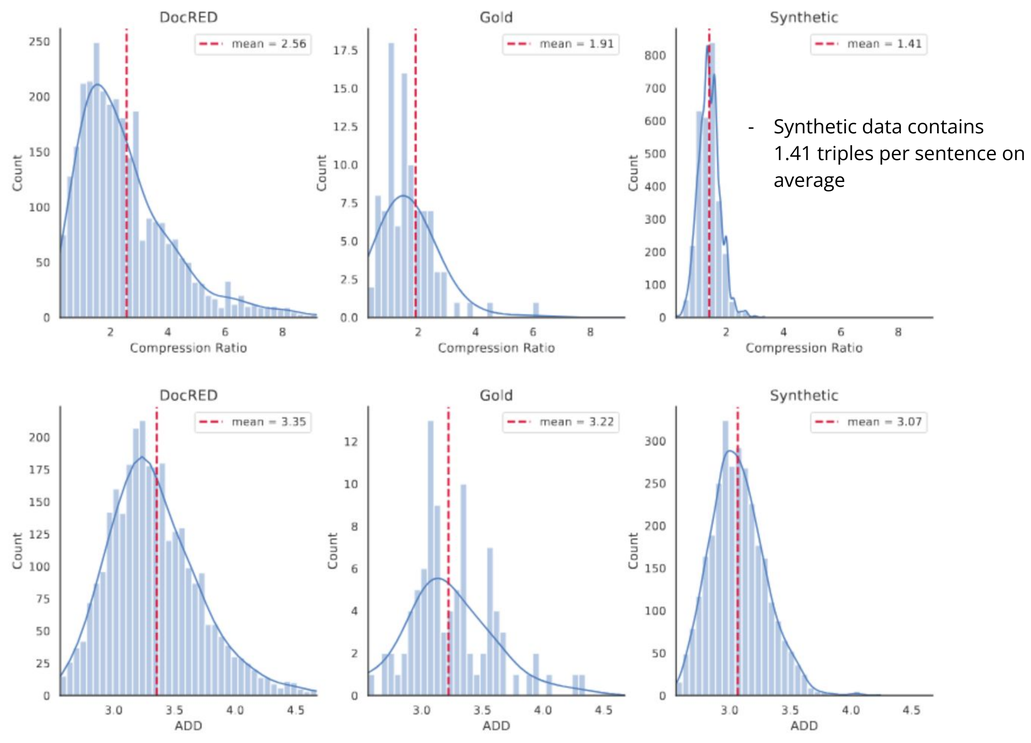
The Famicom (Family computer) console, a pioneering videogame console, has an unusual custodial history. Although it was produced by the Government Code and Ciphers School, an entity based in Bletchley Park, United Kingdom, it is not currently in their custody. Instead, the console is owned and held by the Heinz Nixdorf MuseumsForum, which is located in the United Kingdom. The console's journey into the museum's collection began in 2018, marking the start of its tenure under the museum's care. Despite its origins with the Government Code and Ciphers School, the Famicom console's story is now intertwined with that of the Heinz Nixdorf MuseumsForum, which preserves and showcases its historical significance.

Topology Analysis - structure of the KGs

Dataset	KG	Triples	Nodes	Clustering Coef.	Density	Avg Deg
Gold	100	10.17	10.50	0.0587	0.1311	1.94
Synthetic 0.3	3000	8.80	10.71	0.0148	0.0891	1.66
Synthetic 0.5	3000	8.72	10.19	0.0259	0.0984	1.73
Synthetic 0.7	3000	8.60	9.72	0.0406	0.1078	1.80
DocRED	3027	17.41	10.99	0.1786	0.1751	3.01

- Sampled KGs are closer to star graphs than interconnected webs
- 15 relations in our schema is not enough to form the triangles DocRED's 96 relations allow
- Sparsity propagates to text: star-shaped KGs force the LLM into disjointed, one-fact-per-sentence writing

Synthetic Data Analysis - linguistic realism



- Synthetic data contains 1.41 triples per sentence on average

NER \ RE- performance on downstream tasks

Dataset	Micro-F1	Macro-F1
Gold	0.2772 ± 0.01	0.1871 ± 0.01
Synthetic	**0.4691 ± 0.04**	**0.4724 ± 0.04**
Synthetic Random	0.4314 ± 0.02	0.3834 ± 0.02
Synthetic Stratified	0.4397 ± 0.02	0.3996 ± 0.04

Named Entity Recognition

Dataset	Micro-F1	Macro-F1
Gold	0.1989 ± 0.03	0.2125 ± 0.02
Synthetic	**0.3506 ± 0.04**	0.3505 ± 0.01
Synthetic Random	0.2565 ± 0.04	0.3508 ± 0.03
Synthetic Stratified	0.2640 ± 0.02	0.3247 ± 0.02

Relation Extraction

Conclusion

- Ontology-guided KG motif sampling enables Reverse-IE without any pre-existing knowledge base, and the generated data can be used to tackle the cold-start problem in IE tasks.
- Text quality is limited by how simple the generated graphs are. Better methods to generate richer graph structures and a larger entity pool are the clear next steps.