

Turning legacy data into Linked Data

Kate Byrne

<http://homepages.inf.ed.ac.uk/kbyrne3/>

The Data Silos Problem

Large data collections are often a mixture of relational tables, text documents and multi-media files. Extracting coherent information from related but separate collections is hard—see Fig. 1. Furthermore, whilst the basic information is in easily searchable databases, huge amounts of content are locked in text that is difficult to query systematically.

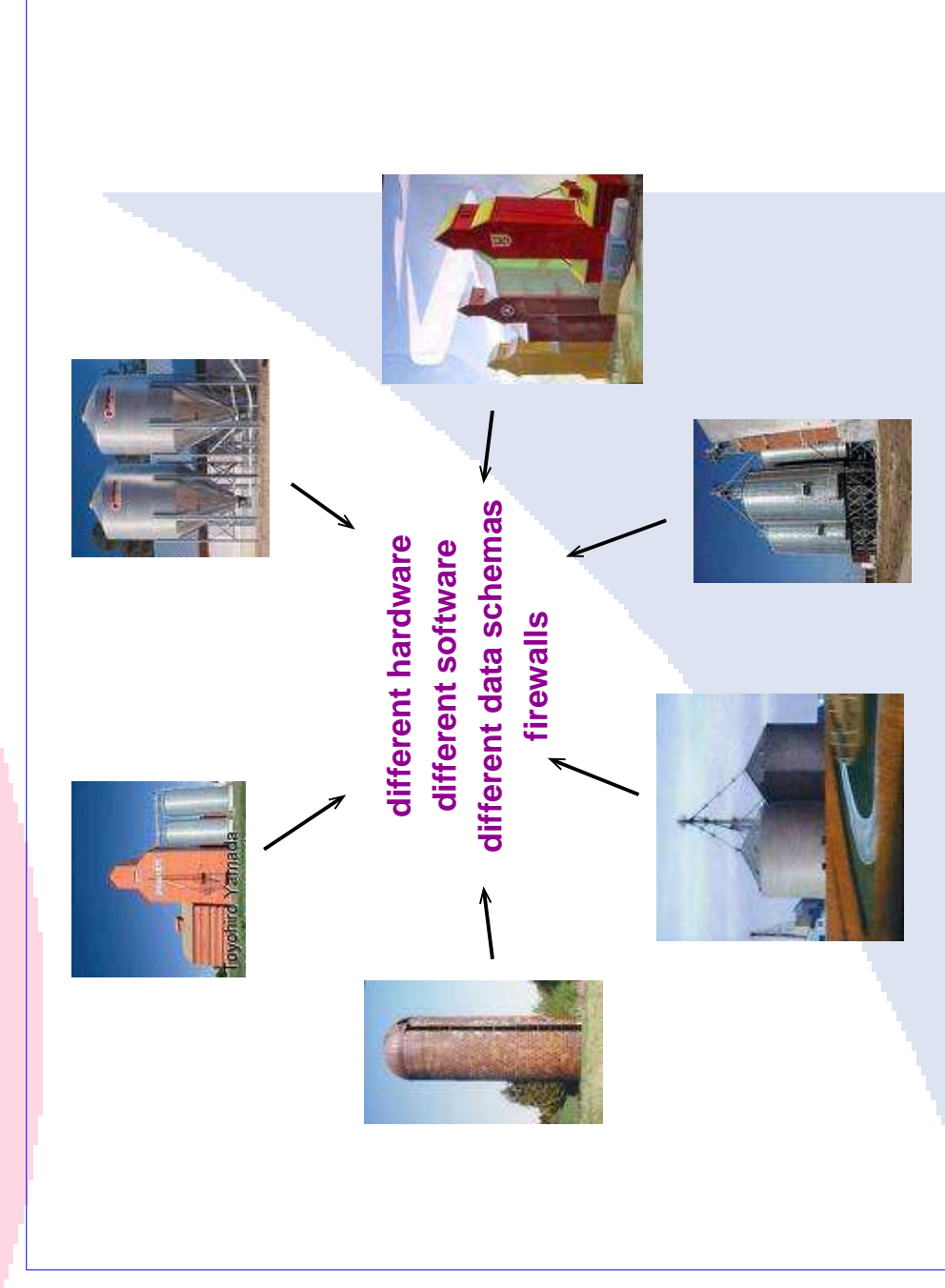


Figure 1: Independent data “silos” are hard to connect.

The semantic web, or “linked data” web, offers a solution. There is a single format—an RDF graph—and bits of data can point to each other as easily as HTML links connect pages on the present web.

Unlocking Text Content

Natural language is our preferred method of communication and most human knowledge resides in text documents. Yet the lack of structure makes text difficult to search effectively and hard to integrate with other data. The real value of the information is lost. The *txt2rdf* pipeline shown in Fig. 2 aims to capture the essential content by extracting factual statements automatically.

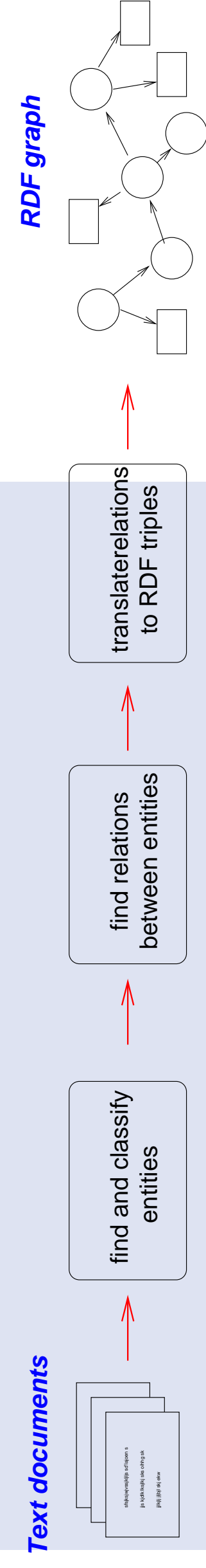


Figure 2: Simplified outline of the *txt2rdf* pipeline.

Firstly there is a named entity recognition (NER) step to find and classify terms like dates and names of people and places. Then comes relation extraction (RE) over the classified entities. Both steps use machine learning, with a model trained on hand-annotated data. Unwanted relations are discarded and the rest are converted to RDF triples. These *Subject-Property-Object* triples are the building blocks of the semantic web.

Figure 3 shows an example, using data from RCAHMS (The Royal Commission on the Ancient and Historical Monuments of Scotland, <http://www.rcahms.gov.uk/>) dealing with archaeological sites.

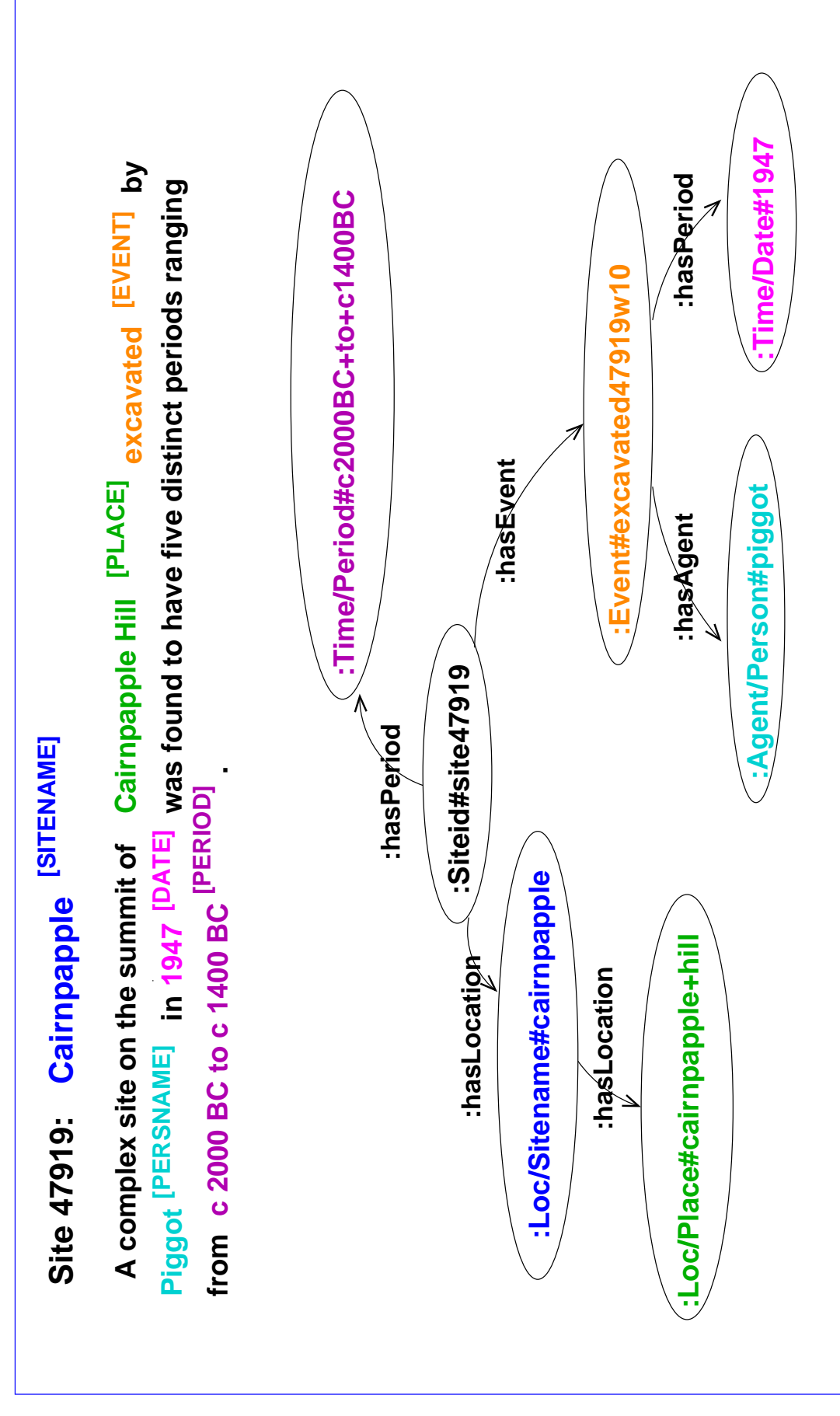


Figure 3: Text relations become RDF triples.

Linked Data

Data from almost any source can be combined using RDF. In this research (Byrne 2009) cultural heritage material from different sources was integrated. The key elements of the system (named *tether*) are shown in Fig. 4, including the *txt2rdf* pipeline.

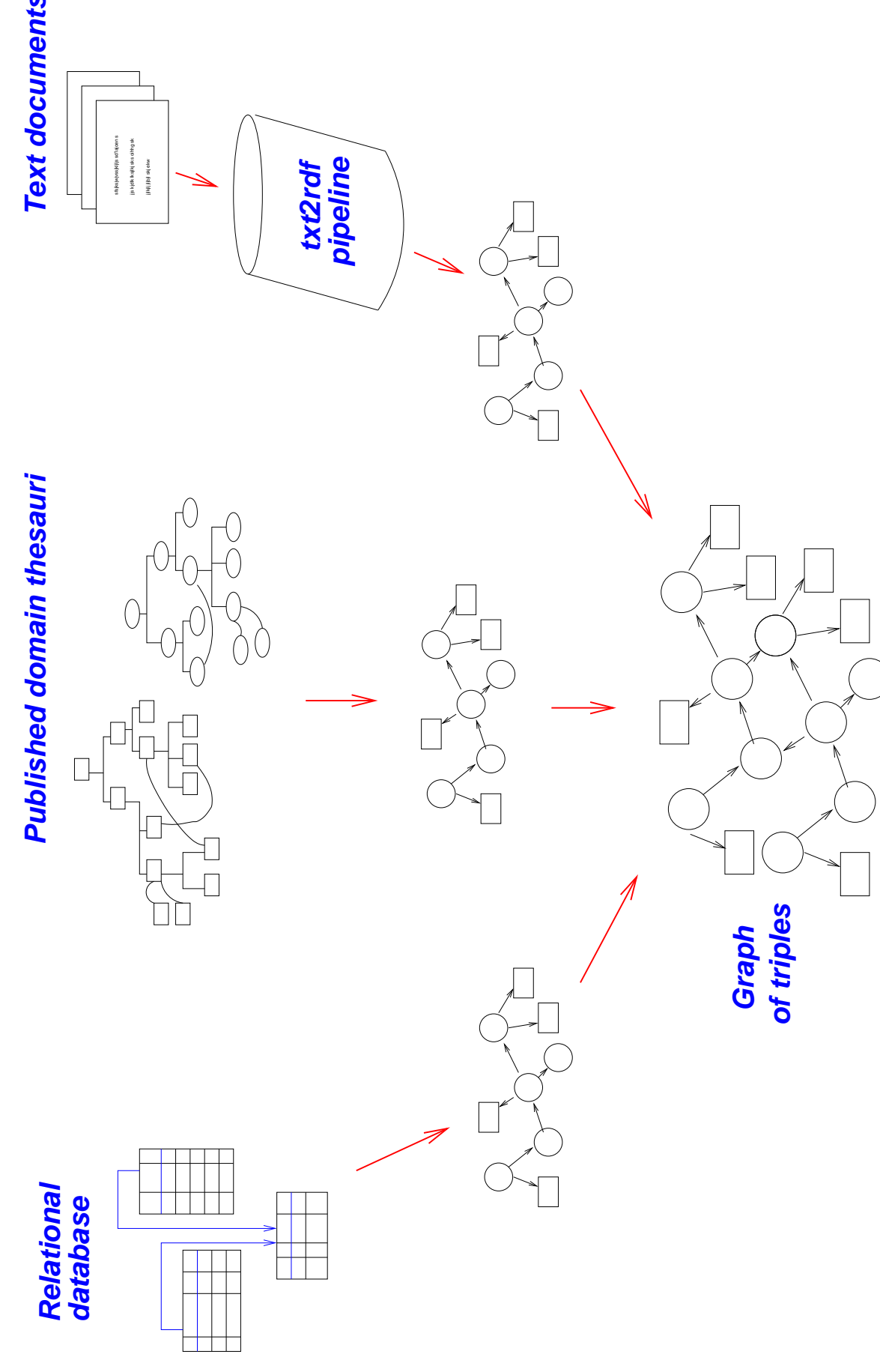


Figure 4: An overview of the *tether* System.

Transforming the database content to RDF means that it can be easily combined with the facts extracted from text documents, with grounding data from ontologies and gazetteers and, most importantly of all, with other RDF data held anywhere in the world.



If data from separate “silos” is transformed in this way, the connections may arise automatically—if there are common nodes, the graphs are already linked. Figure 5 shows an example from the cultural heritage domain: connecting archaeological site data to museum finds via a shared site name (the highlighted node).

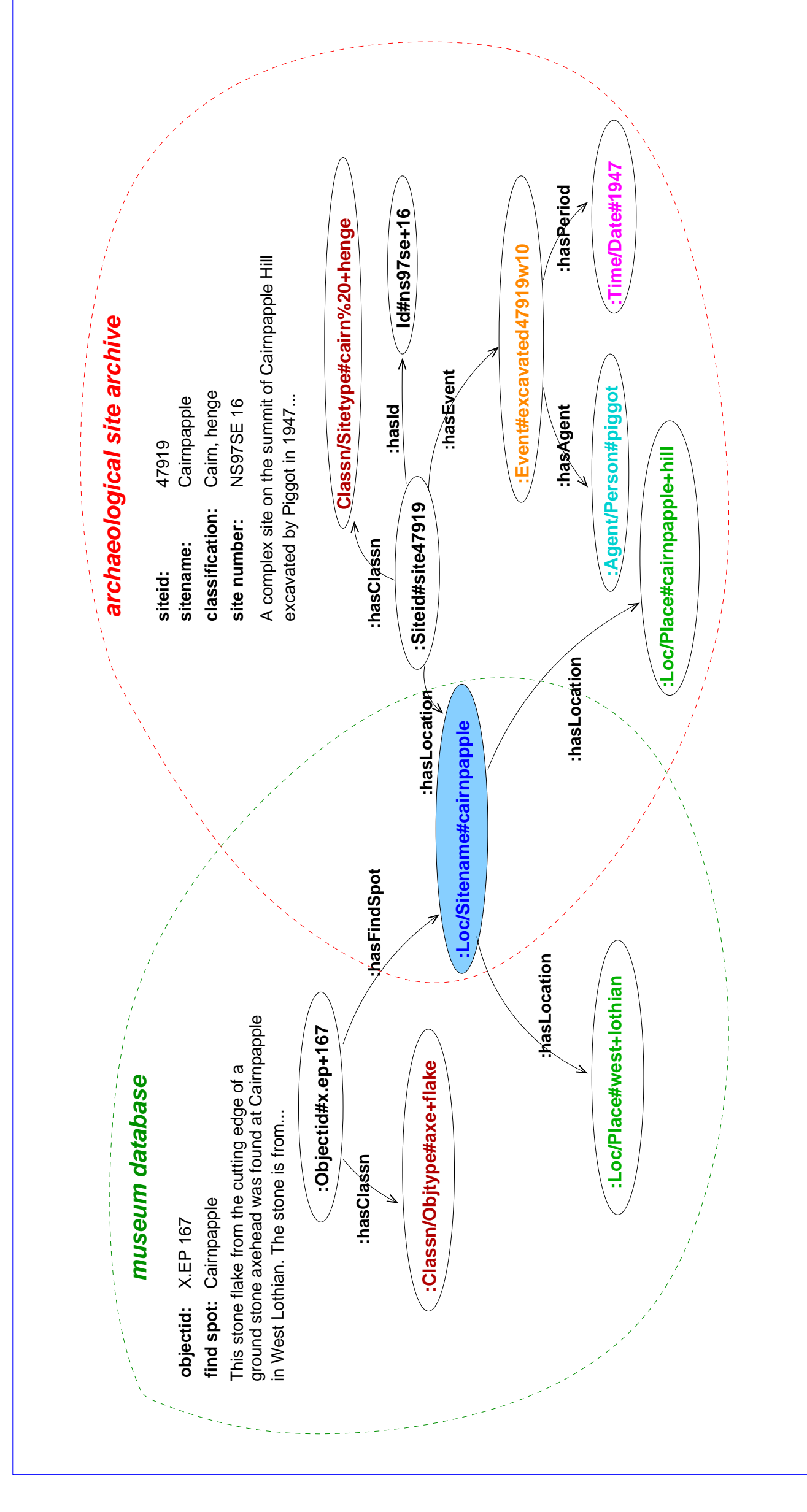


Figure 5: Linking data collections using RDF.

If the graphs don’t connect automatically an extra step is needed, to insert links between related nodes in each. Adding new links in RDF is very easy; the tricky part is working out where to place the links. Trying to make valid connections automatically, or with limited human input, is an ongoing research problem.

Applications and Opportunities

The flexibility of the RDF data structure means there are any number of possible applications of the technology:

- connecting up the silos
- semantic search engines
- reasoning over the data by intelligent agents
- re-purposing data, eg natural language generation from RDF triples
- populating traditional databases by mining text.

References

Byrne, K. (2009). *Populating the Semantic Web—Combining Text and Relational Databases as RDF Graphs* PhD thesis, University of Edinburgh.