

Proposed Annotation for Entities and Relations in RCAHMS Data

Kate Byrne

3 Dec 2006

1 Introduction

A collection of 1546 documents from the RCAHMS dataset was annotated with Named Entities (NEs) for the SEER project in 2003. The document files are a randomly chosen subset of the text notes fields from the RCAHMS database; one text field per file. They vary in length from a couple of lines to about a page. The text is in the form of rough notes: only a minority of the snippets form grammatical English sentences. See Nissim and Krymolowski (2003) for a description of the entity classes and subclasses used. Some familiarity with these earlier guidelines is assumed here.

It is proposed that further annotation is now added to this data to correct some errors, alter entity classes and add relations. Apart from the error correction, new annotation should supplement rather than replace the existing. This document details the proposed changes.

The existing NE annotation is described in Section 2, along with proposals to alter it to fix errors and check the pre-processing steps such as tokenisation. Much of this work can be done automatically by writing suitable programs. These changes would involve revising the existing corpus. Section 3 describes a proposed new layer of NE annotation, which would become available as an alternative NE structure. Some of this layer can be built automatically but most of it will require human judgment. Section 4 covers the proposed relations annotation.

2 Existing Annotation

2.1 Current NE classes

The 1546 files that were originally annotated for SEER are currently held under the directory `/group/ltg/projects/SEER/Data/RCAHMS`. There are various subdirectories there, holding different versions of the corpus. The *StandOff* directory contains the original data, as pairs of files, e.g. `000101.words.xml` for the tokenised and POS-tagged source text and `000101.entities.xml` for a set of pointers to entities within the text. The *knitted* directory contains the same data knitted into single files with inline markup.

The markup is not exactly as described in the guidelines document (Nissim and Krymolowski, 2003); the tags actually used are shown in Table 1. The counts in the table are of entity occurrences, each entity being made up of one or more tokens. Every NE was assigned to a **class**, and in some cases the class was further divided into **types** and **subtypes**. Multiple type tags were permitted (e.g. “industrial, social”).

Up to three levels of entity nesting were allowed. For example,

```
[[[Edinburgh]PLACE University]ORG Library]ORG
```

has three levels. The counts in Table 1 are the totals across all levels of nesting, so `Edinburgh University Library` would be counted three times, for the PLACE entity and the two ORG ones.

2.2 Proposed revisions

2.2.1 Removing Subtypes

The original guidelines nominated three Types under the PLACE class: “administrative”, “natural object” and “specific location”, with the “administrative” Type being divided into Subtypes. However, this is the *only* use of Subtypes (see Table 1) and many entities are already marked without the extra division; so the following changes are proposed, to standardise the markup and eliminate Subtypes altogether:

| Current | New |
|-----------------------------------|------------------|
| PLACE – administrative – council | PLACE – council |
| PLACE – administrative – country | PLACE – country |
| PLACE – administrative – district | PLACE – district |
| PLACE – administrative – parish | PLACE – parish |
| PLACE – administrative – region | PLACE – region |

This change can be automated with very little programming effort.

2.2.2 Fixing case errors

All Type labels will be converted to lower case. This can be automated with very little programming effort.

2.2.3 Marking of spaces

For reasons connected with the functionality of the annotation software, the spaces between tokens were originally converted to “word” elements in their own right, e.g.

```
<W id="w-1-3" C='SPC'> </W>
```

This is non-standard and may cause difficulties further down the pipeline (for example with the chunker), so these elements are to be eliminated and replaced with a single space. This can be automated with very little programming effort.

| Entity Count | Class | Type | Subtype |
|--------------|-----------|---|----------|
| 1869 | LINK | | |
| 2605 | ORG | | |
| 370 | PERIOD | | |
| 2273 | PERSNAME | | |
| 588 | PLACE | administrative | council |
| 103 | PLACE | administrative | country |
| 5 | PLACE | administrative | district |
| 803 | PLACE | administrative | parish |
| 27 | PLACE | administrative | region |
| 36 | PLACE | council | |
| 147 | PLACE | district | |
| 1042 | PLACE | natural object | |
| 44 | PLACE | parish | |
| 31 | PLACE | region | |
| 2 | PLACE | ship | |
| 2072 | PLACE | specific location | |
| 1 | PLACE | terrestrial | |
| 67 | PLACE | transport | |
| 1797 | REFERENCE | | |
| 2 | SITE | | |
| 90 | SITE | industrial | |
| 69 | SITE | INDUSTRIAL | |
| 2 | SITE | industrial,social | |
| 6 | SITE | industrial,terrestrial | |
| 4 | SITE | monument | |
| 228 | SITE | religious | |
| 351 | SITE | RELIGIOUS | |
| 40 | SITE | religious,terrestrial | |
| 232 | SITE | social | |
| 586 | SITE | SOCIAL | |
| 80 | SITE | social,terrestrial | |
| 2 | SITE | social,transportation and infrastructure | |
| 331 | SITE | terrestrial | |
| 873 | SITE | TERRESTRIAL | |
| 2 | SITE | terrestrial,transportation and infrastructure | |
| 154 | SITE | transportation and infrastructure | |
| 363 | SITE | TRANSPORTATION AND INFRASTRUCTURE | |
| 12 | SITE | unassigned | |
| 671 | SITE | UNASSIGNED | |
| 65 | SITE | unspecified | |
| 2185 | SIZE | | |
| 4 | TIMEX | DATE | |
| 3540 | TIMEX | DATE | |
| 1 | TIMEX | TIME | |

Table 1: SEER annotation

2.2.4 Pre-processing steps

It's not completely clear how the original tokenisation and POS tagging was done, and it is suggested that it is checked. This will probably be at least a day's work; more if the tokenisation or tagging needs changing.

3 New NE layer

The aim is to tidy and simplify the existing NE structure, and to introduce new distinctions within NE classes. A new layer would be added to the files so that one could choose to use either the original or new NE structure.

The new entity classes are as described below, with an indication of how they relate to the original classes. Where possible the additions will be made automatically, but where an existing class is to be split between two or more new classes the alterations will have to be done manually. The files presented for annotation will have the new annotation generated as far as possible. The automatic classifications should be checked as part of the annotation exercise.

Each entity is to be tagged with a class tag, from the following list: ORG, PERSNAME, ROLE, SITETYPE, ARTEFACT, PLACE, SITENAME, ADDRESS, PERIOD, DATE, EVENT. For SITETYPE, EVENT and ARTEFACT, there will be a further division into subclasses, as described in Section 3.1 below. The SIZE class is dropped. The LINK class is dropped, as co-reference will be treated as a relation between entities, as described in Section 4. The REFERENCE class is dropped, because bibliographic reference will be treated as a DESCRIPTION event, with an associated eventRel relation (see Sections 3.12 and 4.7).

Entity nesting will be retained unchanged, with a maximum of three levels as before.

3.1 Instances and Classes

Each text string that is marked as a Named Entity will automatically be given a unique identifier by the annotation software. For all entity classes except SITETYPE and EVENT this is all that is required, as the members are distinct individual instances, such as "Mr. J.D. Jamieson" (a member of the set of PERSNAMEs) or "15 May 1968" (a DATE).

The members of SITETYPE will be strings like "chambered cairn", which may be generic or specific:¹ compare "the Dwarfie Stane is a chambered cairn" with "the chambered cairn is in Hoy And Graemsay". For the specific case, the unique instance identifier given by the software must be used in any relation marking. In the generic case, "chambered cairn" is the name of a set — of entities that are chambered cairns. Therefore an extra class such as CHAMBERED-CAIRN is needed, which will be a subclass of SITETYPE, and of which the specific instance.(say sitetype-123) is a member. Figure 3.1 illustrates the point.

¹To think of it in terms of relational algebra, the strings associated with SITETYPEs sometimes represent variables and sometimes values. In contrast, PERSNAMEs (for instance) are always values, of type PERSNAME.

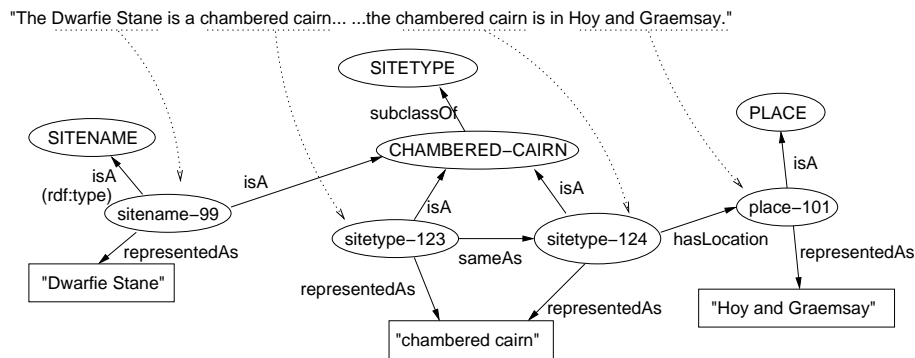


Figure 1: Example of relations from text, showing introduction of CHAMBERED-CAIRN class as subclass of SITETYPE.

The practical implication is that each entity classified as SITETYPE must have a subclass identified. These will be taken from the Thesaurus of Monument Types. There are several hundred terms, too many to list here.

Much the same applies to EVENT entities, but the set of subclasses is more manageable: SURVEY, EXCAVATION, FIND, VISIT, DESCRIPTION, CREATION and ALTERATION. Each text string marked as an EVENT will get an identifier (say event-456) and will be a member of one of the EVENT subclasses, say VISIT. See Section 3.12 for further details.

ARTEFACT entities will also be sub-categorised, as described in Section 3.6 below. However, these NEs are not expected to act as the range for relations — they will be used as instance terms not class terms — so they are less complex than SITETYPE to deal with, and the categorisation is just for convenience.

3.2 ORG

This is the original ORG class, more or less unchanged. It covers organisation names such as “RCAHMS”, “Ordnance Survey”, “OS” etc. It should also be used for named archive collections (usually identified in the text by the word “Collection”) or architectural practices, such as “Walker & Duncan”. In cases like this the component entities will also be tagged (as nested entities), typically as PERSNAME.

3.3 PERSNAME

This is the original PERSNAME class, unchanged. It covers personal names and initials representing individuals, such as “Vere Gordon Childe”, “Henshall”, “RJB” etc.

3.4 ROLE

This is a new class, not corresponding to any of the original ones. It is for the rôles of agents (usually PERSNAMEs), where these are specifically mentioned, for example

“architect”, “donor”, “Chief Executive”. In the archaeological data it will occur only rarely.

3.5 SITETYPE

This is a new label. It will include site classification terms like “chambered cairn”, “long barrow”, “cist” and so forth. Previously these would have come under SITE, which has now been split into SITETYPE and SITENAME. All of the NEs previously categorised as SITE, with a Type label from TMT (Thesaurus of Monument Types), such as “industrial”, “religious”, “social”, will now be classified as one of the subclasses of SITETYPE (as explained in Section 3.1).

Very often the text string will match an entry in TMT exactly, but a more general phrase may be marked when it is clearly describing the type of site. In these cases a best guess at the nearest TMT entry should be made, to serve as the subclass name.

3.6 ARTEFACT

This is a new class and does not correspond to any of the old classes. It is intended to cover terms denoting physical objects that have a significant relationship with the parent site described by the document, yet are distinct from it rather than an integral part. This would include archaeological finds such as “bronze axe”, “sword”, “pottery shards” and so on. Despite the name, it is not restricted to man-made objects, but should include items such as “human remains”, “dog bones”, “seeds”, etc. It should not be used for objects that are components of the parent site, such as “the second sheiling” or “this mound”. As a general rule ARTEFACTs will be portable items that could in principle be separated from the parent site without losing their identity.

Each ARTEFACT entity should be given a class label from the thesaurus supplied, in the same way that SITETYPES are subclassified. If the text string does not match any entry exactly, the best available label should be picked.

There will be some cases where it is not obvious which of SITETYPE and ARTEFACT is more appropriate — where there is a description of inhumations (human burials) for example. If the parent site is a burial site, and the NE term refers to, say, a burial chamber which is part of the site, SITETYPE should be used. Where the term refers to the remains contained in a burial then use ARTEFACT. There will be cases where a best guess must be made.

3.7 PLACE

In the original annotation, PLACE was divided into several Types, some of which had Subtypes. Most of the existing PLACE NEs will be transferred into the new PLACE class, but some will need to be reclassified as SITENAME or ADDRESS. PLACE should be used for all of the administrative place names: regions, districts, parishes, counties and countries, for example. Most of the old “natural object” Type will also come under PLACE. It is intended to include the kind of names that might appear on a

map,² such as “Dumfries and Galloway”, “River North Esk”, “Arthur’s Seat”.

3.8 SITENAME

This is a new class, whose members will come from some of the old PLACE and SITE classes. It should be used for terms that name a specific site, such as “Skara Brae”, “Maes Howe”, “Stones of Stenness”.

3.9 ADDRESS

This is another new class, intended to pick up terms that describe the location of a site, rather than simply naming it. Part of the aim is to avoid cluttering the PLACE class with very local terms that are unlikely to be useful for querying the final graph. For example, architecture texts often include NE terms such as street names, or even postal addresses. In archaeological texts grid references are common, such as “HU 3754 3380” and site references like “HU33SE 43”,³ and both should be marked as ADDRESS. This label may be appropriate for some of the existing “specific location” NEs. Private places like “John’s farm” would become members of the ADDRESS class (unless they indicate a specific archaeological site, when SITENAME should be used). Names that are too local for the “map rule” described for the PLACE class — such as street names or house names — should be classified as ADDRESS.

3.10 PERIOD

The existing PERIOD class remains unchanged. It covers terms such as “late Neolithic”, “16th Century”, “modern” and so on. Terms identifying a particular calendar date come under DATE.

3.11 DATE

The existing TIMEX class is replaced by DATE, and now only covers date values, such as “1st Jan 1980”, “October 2002”, etc. Existing members of the DATE Type within TIMEX are to be transferred into the new DATE class, but non-specific time references like “Sunday morning” or “the following week” are not required. The TIME Type, which was a subset of TIMEX, is no longer required.

3.12 EVENT classes

This is a new family of classes, not corresponding to any of the old classes. It will cover terms describing events in the history of a site, such as visits, surveys and excavations. The subclasses of EVENT are: SURVEY, EXCAVATION, FIND, VISIT,

²This means a map intended for the general user — say a walker or car-driver — not a specialist map of archaeological sites, nor a detailed large scale plan. The OS 1:50,000 map would be a typical example.

³“HU33SE” is the number of a particular OS map sheet, and “43” is the number allocated by the NMRS for a particular site on that sheet. The combination constitutes a unique identifier for the site. There is often a sub-number as well, expressed as “43.1” or “43-1” or similar.

DESCRIPTION, CREATION and ALTERATION. The first five event types listed are expected to be the most frequently encountered in this corpus, but the last two (creation and alteration) can be used when appropriate.

The EVENT classes all pertain to n -ary relations that are awkward to translate into simple two-place predicates. Each event will typically have a subject or “patient” (generally the site itself), an agent (the instigator of the event) a date, and possibly other attributes or semantic roles (such as what was found, in a FIND event). These linked NEs will be picked up through the relation annotation, as described in Section 4 below. The event entities are needed to provide a hook to hang the relation annotation on.

In many cases the events are implied by the text, not explicitly mentioned. For example, phrases like “visited by OS, 1990” are common; clearly a visit took place, but there is no noun phrase to label. In such cases the verb phrase (“visited”, in this case) should be marked as a VISIT event entity. In other cases a suitable nominal will be present, as in “...excavation carried out by...”, where “excavation” is an EXCAVATION event.

In more detail, the different subclasses cover events as follows:

1. *SURVEY*: This is the appropriate category when the text contains references to “plan”, “survey”, “measured survey”, “GDM survey”, “photographic record” or such like. It implies a detailed examination of the site resulting in the production of archive material.
2. *EXCAVATION*: Self-explanatory - there will be a reference to an excavation or dig, i.e. an event in which the site was deliberately physically disturbed by an archaeologist.
3. *FIND*: When an ARTEFACT entity occurs, there should typically be an associated FIND-EVENT, which will generally be expressed as a verb phrase such as “was discovered”, “found” etc.
4. *VISIT*: This is a fairly unspecific event, when an organisation or person went to the site but there is no reference to a survey or excavation.
5. *DESCRIPTION*: This is the most general category, used when it’s not clear that the site was visited at all, but some agent is mentioned as having produced a tangible description or depiction. It is also to be used for bibliographic references, as follows. Where there is a suitable noun or verb phrase (as in “...described by E Beveridge”) this should be marked as the DESCRIPTION NE (“described” in this case). Where there is only a free-standing bibliographic reference (such as “E Beveridge 1911”) the whole string should be marked as a DESCRIPTION, with nested entities inside it (typically PERSNAME or ORG, and DATE).⁴
6. *CREATION*: This category, and the next one, will be uncommon in the RC-AHMS archaeological data that comprises most of the SEER corpus, but is included to provide coverage for architectural data. A CREATION event refers to the original construction of a monument. In the case of a building, several agents may be mentioned (architect, builder, patron, etc.) as well as a date. If there are multiple rôles or dates, separate EVENTS should be used. This may necessitate relabeling a string already marked as a CREATION event. (See also Section

⁴The aim is to avoid nesting entities if possible, but some text string always needs to be marked as the NE that takes part in an eventRel relation.

4.7 and Figure 3.) As for DESCRIPTION, where there is no obvious noun or verb phrase (such as “built by”) available, a suitable string must be identified for labeling, usually the compound of the entities participating in the event (e.g. “Architect: William Adam 1723”).

7. *ALTERATION*: This is intended to cover events that change the physical character of a site significantly, such as serious damage, extension, transfer of location, etc. Occasionally a monument that no longer exists is recorded, and there may be information on when it was destroyed. This should also be classed as an alteration (the cases are too rare to warrant a separate DESTRUCTION class).

4 Relation Annotation

The relation annotation will cover the new NE layer described in Section 3. With one exception, the relations will be (subject, predicate, object) triples,⁵ where the subjects and objects are NEs. The predicates are: *isA*, *sameAs*, *seeAlso*, *partOf*, *hasLocation*, *hasPeriod*, *eventRel*. All except the last mentioned are triples as just described, but the *eventRel* relation has higher arity and will be of the form: *eventRel*(*eventType*, *eventPatient*, *eventDate*, *eventAgent*, *eventRole*, *eventResult*). The *eventRel* relations will subsequently be transformed into binary relations, but the proposed arrangement is intended to make the annotation process simpler.

The characteristics of the relations are as described below. Where possible relation names from published ontologies will be used (such as *rdf:type*, *owl:sameAs*), but the renaming will be done in a later processing step and is not detailed here. The most common domain and range of each relation is given for guidance, but there may be cases where they do not apply. In linguistic terms the domain of a binary relation is typically the subject or agent in a textual expression, and the range is the object or patient.

4.1 *isA*

Domain/Agent: SITENAME

Range/Patient: subclasses of SITETYPE

Example: “Hill of Caldback is a chambered cairn”

This indicates membership of a class, but there is no need to add an *isA* relation to show the class of every single entity, such as (“RCAHMS”, *isA*, *ORG*), as this will be done automatically. The *isA* relation is intended here to indicate multiple inheritance, where an instance belongs to another class as well as its “natural” parent. In this domain, this is most likely to occur with membership of the SITETYPE subclasses. Wherever possible the specific subclass should be used (such as *CHAMBERED-CAIRN*) rather than the parent SITETYPE class.

4.2 *sameAs*

Domain/Agent: any NE instance

⁵- or, equivalently, two-place relations of the form *predicate*(*subject*, *object*)

Range/Patient: NE instance of the same class
Example: "...described by A. S. Henshall, 1985. Henshall also says..."

Use this for co-reference. In the original SEER annotation the LINK class was used, and only pronouns and definite descriptions were marked. For this new annotation layer, *all* co-referential mentions of entities should be included, as the example illustrates. Any number of entities can be marked as belonging to the same co-reference set. They will typically be entities likely to feature as nodes in the final graph, i.e. as the source or target of a binary relation. Hence relative pronouns, which were previously marked as LINKs, are not required.

Note that the relation applies to *instances* not classes. For the SITETYPE class, NE strings such as "chambered cairn" in "Hill of Caldback is a chambered cairn" will have a unique label assigned, such as sitetype-123, as has already been discussed (in Section 3.1). If the same entity is referred to elsewhere in the text as a "chambered round cairn" (perhaps sitetype-456) the two SITETYPE instances should be linked by a sameAs relation. This does not imply that in general a chambered cairn is identical to a chambered round cairn, because those are class terms and the two classes are not equivalent.

4.3 seeAlso

Domain/Agent: any NE instance (often SITENAME or ADDRESS)
Range/Patient: usually an NE instance of the same class
Example: "See also HU33SE 43 excavated by Parry"

This is intended for occasions when two entities are described as being closely related, or are contrasted with each other. It may also be used when it seems sensible to link two entities, but the relationship between them is not in the specified set, such as a family tie (parent, child, sibling, etc.) between two PERSNAME entities, or some noteworthy association between, say, a SITENAME and a PERSNAME (such as "Sir Walter Scott lived here"). The relation is bi-directional, so the order of subject and object is not significant. However, for examples like the one given above, the convention to follow is that the current site (the subject of the text) is the subject and the entity pointed at (in this case the ADDRESS, "HU33SE 43") is the object.

4.4 partOf

Domain/Agent: SITETYPE
Range/Patient: SITETYPE or SITENAME
Example: "A farmstead comprising one unroofed building, two roofed buildings and one enclosure, and a head-dyke"

This is for part-whole relationships, such as when a complex site is described in terms of its components. In the example given, the "farmstead" SITETYPE is the whole, and "unroofed building" "roofed buildings", "enclosure" and "head-dyke" are the parts. Each of them is a SITETYPE NE.

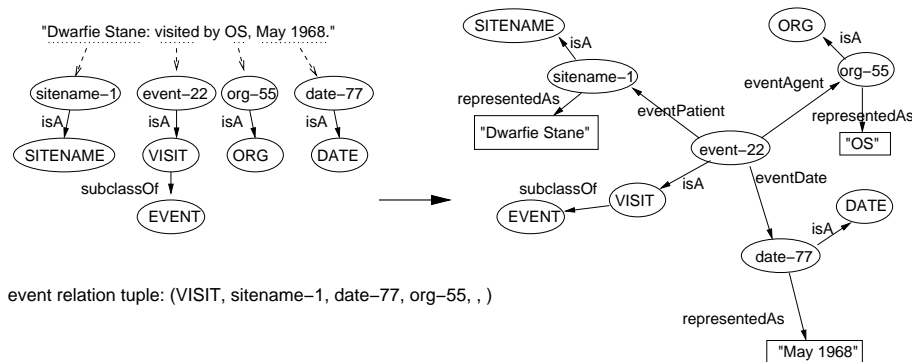


Figure 2: Example event relation tuple, and its translation to RDF graph format.

4.5 hasLocation

Domain/Agent: SITENAME, SITETYPE, ORG, PERSNAME, ARTEFACT
 Range/Patient: usually PLACE or ADDRESS; may be ORG (see below)
 Example: "...on the north side of the road leading to Bannaminn"

This relation covers cases where an entity is located at or in the vicinity of a PLACE or ADDRESS. In the case of rather vague descriptions like the one in the example, the consideration should be whether knowing the location mentioned would help someone to find the entity. (In this case it would.) Negative examples (if they occur), such as "a long way from Edinburgh" should therefore be ignored. For ORGs and PERSNAMEs the relationship will typically be with their addresses; for ARTEFACTs it will often be with the museum holding them. If a PERSNAME is mentioned as belonging to an ORG, this should be marked as hasLocation.

4.6 hasPeriod

Domain/Agent: SITENAME, ARTEFACT
 Range/Patient: PERIOD
 Example: "a late Viking potsherd"

This is a straightforward link to PERIOD NEs from the NE they apply to.

4.7 eventRel

Event relations connect a set of entities taking part in one of the events defined as subclasses of EVENT in Section 3.12. They will be converted to RDF binary relations in a later processing step. The tuple making up the relation is of the form: (eventType, eventPatient, eventDate, eventAgent, eventRole, eventResult).

Figure 2 shows the tuple for the example text "Dwarfie Stane: visited by OS, May 1968", and how it can subsequently be translated into a graph of two-place relations.

In more detail, the arguments are:

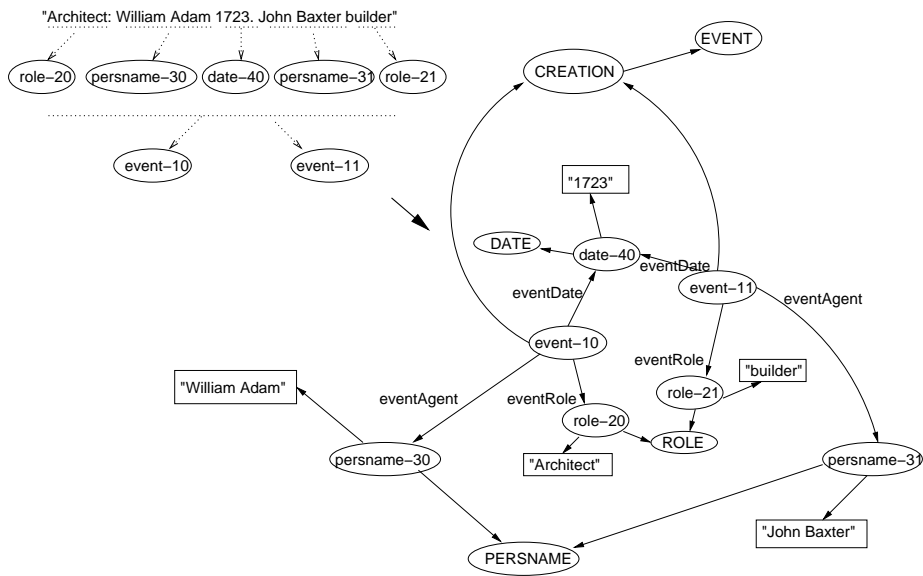


Figure 3: A CREATION event, split into two events related by `sameAs`, to cater for multiple agents and rôles. (Only the key edge labels are shown.)

1. `eventType`: one of the set {SURVEY, EXCAVATION, FIND, VISIT, DESCRIPTION, CREATION, ALTERATION}; required, and occurs once
2. `eventPatient`: the object undergoing the event, filling the direct object position for active verbs; null if no object mentioned, and can occur only once
3. `eventDate`: the DATE when the event took place; null if no date mentioned, and can occur only once.
4. `eventAgent`: the PERSNAME or ORG that was the instigator of the event; null if no agent mentioned, and can occur only once
5. `eventAgentRole`: the ROLE of the `eventAgent`, where this is explicitly mentioned (e.g. “architect”). In some events, such as CREATION, there may be multiple PERSNAME agents playing different rôles (architect, designer, etc.). In these cases, each `eventAgent`–`eventAgentRole` pair should be put in a separate `eventRel`. This will involve the same text string being labeled as more than one CREATION event NE. See Figure 3 for an illustration. (The aim is to avoid the need for a more complex structure, with new relations, to tie each agent to the correct rôle.)
6. `eventPlace`: where the event took place, if this is obviously part of the relation; it would be useful where some location *other* than the current SITENAME is mentioned; optional and non-repeating

References

Malvina Nissim and Yuval Krymolowski. RCAHMS Annotation Guidelines. Held at `/group/ltg/projects/SEER/Annotation/Guidelines/guide.tex`, May 2003.