https://doi.org/10.2352/issn.2168-3204.2018.1.0.31

Setting out on an unknown sea – an extremely flexible metadata model for the "Engelandvaarders" collection (a case study)

Martijn van der Kaaij; Heron Information Management LLP; Weert; The Netherlands

Abstract

To address the very diverse and still developing requirements of maintaining and managing a growing collection of data on "Engelandvaarders" (people who escaped from the occupied Netherlands to England during the Second World War to continue the fight against the Germans), a flexible data model was proposed, built on semantic triples.

This approach was expected to result in a) an enduring ability to deal with new categories of resources b) a very significant reduction – after initial development - of the need for work on database interfaces, both for data entry and for data viewing and c) creation of a portable, platform independent and application independent dataset. These results were achieved, and in addition it was discovered that the semantic approach notably improved communication on the metadata requirements within a varied group of stakeholders, volunteers and developers.

Finally, visualization benefits were expected, but the actual results surpassed those expectations.

Background

In September 2015 the "Museum Engelandvaarders" [1] opened its doors to the public. The museum, situated in a German bunker dating from the second world war on the Dutch coast in Noordwijk, tells the story of the "Engelandvaarders" (= those who sail to England), predominantly young people who tried to flee the occupied Netherlands in 1940 – 1945 to reach the government in exile in London, with the intention of continuing the battle against the Germans from there.

The museum has limited space, and is therefore only able to display a small part of its collection, and many resources in its collection (mainly images, videos and textual documents) remain hidden from the visitors. Furthermore, resources are still forthcoming: surviving Engelandvaarders or their families keep offering images and documents to the museum. Also, the growing awareness among the public of the existence of the museum has started to generate a steady stream of questions.

In order to make sure those 'hidden' resources would be registered, and those questions could be answered, the museum board initiated the development of a collection database some months before the opening of the museum. This database was expected to serve two purposes:

1) Establish a register of all known Engelandvaarders, with a short summary of their fate.

2) Offer all available resources (images, documents, video etc.) to researchers.

By now, a third purpose has been added:

3) Present narratives about incidents involving Engeland-vaarders. These narratives form a third way of accessing the resources in the database.

At this moment in time, the database is only accessible on two displays in the museum. However, the chosen technology is entirely web based, and could easily be published to the wider world. For the moment, however, the board of the museum prefers those interested to show up at the museum, or to ask for specific information, which will then be collected and sent.

Phase 1: limited data modeling on the fly

The founders of the Museum Engelandvaarders had a dual purpose in mind: the museum should tell the story of the Engelandvaarders by a combination of exhibits and multimedia presentations, but it also should become the national 'knowledge center' on the subject. Having published on information management issues with one of the initiators of the museum, the author was asked to help. It was a matter of boarding a fast moving train. As the database had to be up and running, work on collecting data for the database had already started. Data modeling was mainly driven by 'what can we get?' and not by 'what do we need?' Even so, an initial model was set up. Though none of the people involved would have used the word, two entities were identified: PERSON and (digitized) DOCUMENT.

Apart from the usual attributes regarding name, birth, death and gender, the PERSON had some attributes that were more specific for an Engelandvaarder: faith and occupation before departure were recorded as correlations were expected between certain values for these attributes and the likelihood of becoming an Engelandvaarder. Also, journeys were recorded, stating place and date of departure, place and date of arrival, the chosen route and a note. As the possible routes were limited, a rough characterization was used, e.g. 'southern overland route' (= through France, Spain and Portugal and then by boat or plane to Britain) or 'North Sea'. The note was mainly used to describe the reason for departure (e.g. 'afraid of arrest', 'escape from forced labor'). Furthermore, for those people who made it to Britain, some information could be recorded on the activities after arrival. Finally, one or more sources (literature, archival records) for the information on PERSON were recorded.

Initially, a DOCMENT was modeled in a very simple way: a file name and a caption were deemed sufficient to catch the initial flow of incoming resources.

Data had been collected in Excel files, and was then converted to XML and stored in eXistdb, an open source native XML database [2]. Apart from some minor changes, the database that was revealed on opening day implemented the model set out above.

Phase 2: adapting to (museum) realities

Very soon after opening, the limitations of this initial implementation became evident. Those responsible for the information displays in the museum tired very quickly of the limited format of the register records and asked for the ability to add full blown narratives specifically aimed at presentation on screen in the exhibition spaces. Also, stakeholders on the 'documentation' side were quick to propose refinements of the metadata model to deal with different categories of resources that started to come in.

To be clear: none of this was unexpected. All those involved knew that, as far as the database was concerned, they were setting out on an unknown sea. In order to be in time for

the opening of the museum, the database had to be available in September 2015, but its design was never expected to be finished by that time.

The requests were dealt with: a biographical note was added to contain the narratives, the classification of documents was refined, and the workflow for data entry for images – arriving in bulk at one time – was simplified.

However, the amount of additional requirements and the speed with which they arrived after the opening made it very clear that modifying the existing database on demand would not be a viable option for the longer term. An XML record can, up to a point, publish its own structure through its schema, which means that procedures for data entry and manipulation can be quite generic, but even then, continuous addition of elements and attributes would not be sensible. Would we go down that road, we could expect to work forever on new data entry screens, additional visualizations and storage modifications. Given the limited resources of the museum (almost entirely run by volunteers and operating on a limited budget, of which almost every cent has to be raised by funding campaigns), this was not a viable approach.

Also, while the initial requests for changes mainly came from the museum rooms and were, from a data modeling point of view, quite straightforward, more – and possibly more complex – requests could be expected with the increasing use of the database as a tool for research. Clearly, decisive action was needed.

Phase 3: applying standards and redesigning storage

To address the situation, we started a second phase of data modeling, with the goal of arriving at an ontology for the Engelandvaarders. The ontology should preferably be derived from existing standards, but it should be flexible enough to accommodate new requirements, and it would have to be simple to use, given the skills and interests of the volunteers responsible for data entry.

PERSON

As a first step, we took another look at PERSON. We took inspiration from modeling of persons in the Text Encoding Initiative (TEI) [3], and in particular from the concept of 'Personal Characteristics' (see figure 1).

<faith> specifies the faith, religion, or belief set of a person.
<langKnowledge> (language knowledge) summarizes the state of a person's linguistic knowledge, either as prose or by a list of langKnown elements.
<nationality> contains an informal description of a person's present or past nationality or citizenship.
<sex> specifies the sex of a person.
<age> specifies the age of a person.
<socecStatus> (socio-economic status) contains an informal description of a person's perceived social or economic status.
<persName> (personal name) contains a proper noun or proper-noun phrase referring to a person, possibly including one or more of the person's forenames, surnames, honorifics, added names, etc.
<occupation> contains an informal description of a person's trade, profession or occupation.
<residence> describes a person's present or past places of residence.
<affiliation> contains an informal description of a person's present or past affiliation with some organization, for example an employer or sponsor.
<education> contains a description of the educational experience of a person.
<floruit> contains information about a person's period of activity.
<persona> provides information about one of the personalities identified for a given individual, where an individual has multiple personalities.
<state> contains a description of some status or quality attributed to a person, place, or organization often at some specific time or for a specific date range.
<trait> contains a description of some status or quality attributed to a person, place, or organization typically, but not necessarily, independent of the volition or action of the holder and usually not at some specific time or for a specific date range.

Figure 1. *Personal Characteristics in TEI (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html#NDPERSE)*

These personal characteristics are elements describing physical or socially-constructed characteristics, traits, or states of a person (e.g. social economic status, faith, affiliation, occupation). These are extremely useful in the context of Engelandvaarders, as many assumptions have been made about possible correlations between these characteristics and the decision to become an Engelandvaarder. Recording those characteristics in detail allows for proper research in this area, and will – hopefully – facilitate the step from assumptions to facts. We had been aware of this – faith and occupation were already part of our data model – but this time round we decided to incorporate the full TEI list of Personal Characteristics.

EVENT

While PERSON 'gained' (optional) attributes regarding characteristics, it also lost some other attributes: most information about the things a PERSON does or experiences is now captured by an entity EVENT. Once again, TEI, this time combined with OntoLife [4], provided inspiration.

```
<event xml:id="eMBB" from="1955-12-01"
 to="1956-12-20">
 <label>Montgomery Bus Boycott</label>
 <desc>A political and social protest
campaign against the policy of
   racial segregation on the public transit
system of the city of
 <placeName ref="#MONT">Montgomery</placeName>.
</desc>
</event>
```

Figure 2. *A TEI event (www.tei-c.org/release/doc/tei-p5-doc/en/html/examples-event.html)*

However, some additional modeling was necessary. 'Simple' events at one point in time have a date (precision ranging from year only to a full date), a label (e.g. 'birth'), an optional place and an optional note. If a time span is needed rather than a day – as for journeys and activities after arrival - an end date may be added. To express more complex realities, events may be nested. This last feature is extremely useful to capture the journeys of many Engelandvaarders: their road to Britain often involved many stages, and for each of those stages resources might be available. In this way, we can, for example, link a photo of an airplane to the actual trip from Lisbon to Croydon on a certain date, rather than linking it imprecisely to the full journey from, for example, Gouda to London of which the flight was a part.

RESOURCE

For those resources that we might want to link to a PERSON or an EVENT we have now introduced a RESOURCE entity. This entity replaces DOCUMENT.

If a RESOURCE is an information resource (e.g. texts, images, multimedia) we apply a set of metadata elements based on the 'Information Object' in the CIDOC-CRM model [5]. Otherwise, we consider the RESOURCE, in CIDOC-CRM parlance, to be a 'Man-Made Object' and take our metadata properties (attributes) from there [6]. However, as the CIDOC-CRM is quite a complex mode, we are only using a limited amount of the properties that are part of it. Our main aim here is to be compatible with CIDOC-CRM, if and when it is needed.

As you may have noticed, the inclusion of (museum) objects in our data model is a new feature. The opportunity to link the story of a person not just to digitized documents but also to real objects in the display cases is expected to bring the people in the database to life. It will also allow us to show – albeit only on screen – objects that cannot be displayed in the museum.

NARRATIVE

The most recent extension of the Engelandvaarders ontology is a section for modeling of narratives connected to the Engelandvaarders, which is based on the Curate ontology [7]. Right from the start it has been clear that stories are an extremely important part of the Museum Engelandvaarders. Even in the exhibition spaces, stories displayed on touch screens or projected on the wall compete with the displayed objects for the attention of the visitors. Furthermore, the board of the museum considers it a vital task to capture those stories before it's too late: not many Engelandvaarders remain among the living, and even though some of the families are very involved in the museum, it seems prudent to make haste.

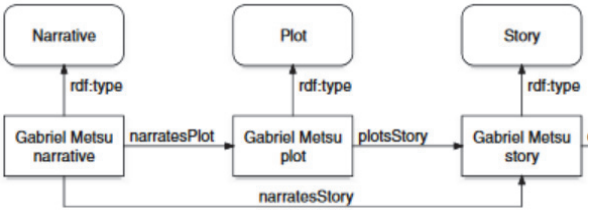

Figure 3. *Relevant concepts of the Curate ontology (https://pdfs.semanticscholar.org/1200/6b4563971515709cf4de82d9a2e05772fb28.pdf)*

In these circumstances, it is important to record the stories, and relate them to people, events, objects and documents as extensively as possible. Curate allows us to do that in such a way that the stories are preserved, while they can be used in the exhibition spaces as well.

While we were investigating Curate, we noticed that, after a flurry of activity in 2012 and 2013, it seems to have gone really quiet around this standard. It could be that all those involved consider the job done, or that some of those involved moved to fresh pastures, but the most logical explanation would be that Curate is not actually used very much. This would reflect the experience of the author that in heritage collections objects and stories are treated as almost unrelated entities. By now, object descriptions are meticulously kept in databases everywhere, but stories about these objects as told in exhibitions or on websites reside in different places, often without a lot of metadata attached. For cultural heritage collections it is essential that those stories are captured, connected to objects and other entities through metadata, and properly preserved.

SOURCE

Above, when we encountered 'source', it was just a repeatable attribute of PERSON. When revisiting our model, it seemed wise to change SOURCE into a separate entity: many of our sources from libraries and archives are now available online, and our SOURCE entity has a repeatable attribute 'identifier', in which the – hopefully persistent – URI of a source can be recorded. As notes can be recorded for a source it allows researchers to share information on – and maybe, in time, even interpretations of – the source in our database.

Authority files

With an eye to, in time, sharing our metadata online, we have made it our policy to relate our data to authority files whenever possible. If, for example, one of our Engelandvaarders is also an author, his identifier from ISNI (International Standard Name Identifier, authority file of choice for Dutch authors) [8] or, if he's not in there, the LCCN (Library of Congress Control Number) [9] is recorded in our database. For places, we use the identifiers from GeoNames [10].

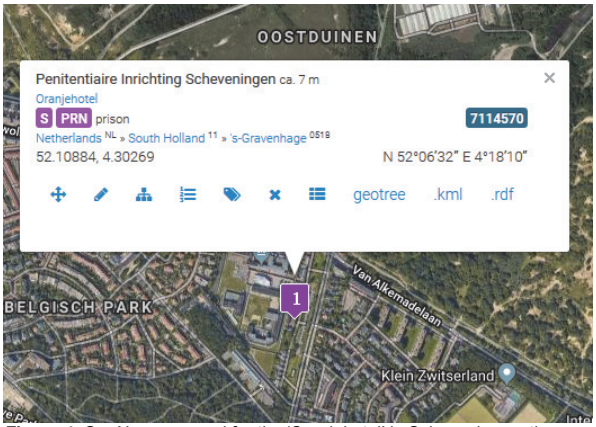

Figure 4. *GeoNames record for the 'Oranjehotel' in Scheveningen, the prison were many Engelandvaarders ended up after being captured by the Germans (http://www.geonames.org/7114570/)*

The examples of authority files above deal with standardization at the level of entities ('is this John Wilkins the same John Wilkins you are talking about? Let's check his identifier'), but we have also included authority files and standard notations at attribute level, for example in using ISO 8601 for date notation [11].

Storage

One of the guiding principles in remodeling the ontology was flexibility: the ontology should be easy to extend with new attributes, new authority files, and even new entities. Our storage application would have to support this flexibility.

Furthermore, we were looking for a solution that would be both platform independent and application independent: in time, the dataset will be one of the treasures of the museum. Whatever happens to the IT architecture or even to the physical museum, the data set must survive. Therefore, it should be entirely independent from data models or (limited) serializations built into applications.

The solution was found in designing the next version of our database as a store of semantic triples. Semantic triples are the atomic data entities in the Resource Description Framework (RDF) data model [12]. A semantic triple is a subject–predicate–object expression, e.g. "Erik Hazelhoff Roelfzema" "has occupation" "student". The advantage of this approach is that, on the storage side, you only need to be able to store triples.

Design of different record structures for different types of data is not necessary. This means that extension of the data model is perfectly possible without the need for extensive software modifications.

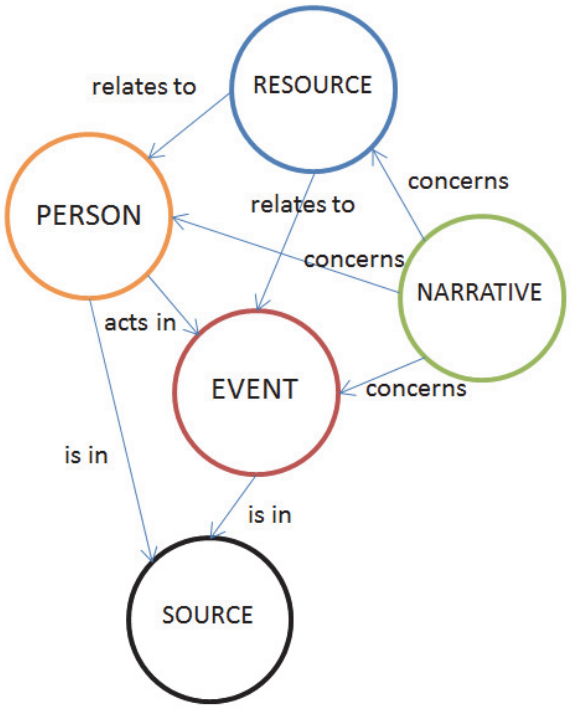

Figure 5. *The Engelandvaarder ontology*

To provide a framework for the definition of our triples – and to be able to be proactive in proposing them – we need, of course, our ontology, as it defines which entities can be related by which predicates. Application of authority files then allows us to replace, at least in some parts of a triple, strings by identifiers, which are far less ambiguous. For our example: <http://www.museumengelandvaarders.nl/p01212> <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-occupation.html> "student@nl", where p012012 is Roelfzema's identifier in the museum database (now the authority file on Engelandvaarders!) and "@nl" indicates that we meant the Dutch word 'student', which, in the absence of an authority file on occupations, at least provides some context.

Our storage model can now be very simple: in terms of a relational database we would only need one table with three columns for object, predicate and subject, instead of tables with field names corresponding to the entities and attributes in our ontology. This means that new predicates and even new entities could be introduced without changing the data structure. Note that this is a functional model for storage. The actual software could still break the triples down and store them in other ways, but that doesn't matter. Any application that allows us to 'talk triple' is now suitable for us. In actual fact, we keep using eXistdb, but now as a triple store.

For those readers who might want to know why we are not using quads [13] instead of triples: because we don't need them (yet), but once we do, they can easily be accommodated in our storage structure.

The chosen approach means that we can easily visualize our data in graphs, showing the connections between all the entities. This has already led to new ways of allowing our visitors to browse through the data. Also, the application of authority files allows us to offer the user of the database meaningful suggestions while he is going through the data.

Taking stock

At the time of writing, we are nearing the end of phase 3: remodeling of the ontology has been completed, and the new entities and attributes are being exposed step by step in the database. Of course, exposing these new features involves converting the existing XML records to triples. These conversions are in progress.

Some data conversion will be necessary as well: stories that have been put into notes for a person will have to be extracted and put out in narratives.

Our next project will be the long term storage of some of our data. Unique objects that are being digitized for our museum should be preserved. To achieve this, we are looking at applying the METS [14] and PREMIS [15] standards.

Conclusions

First and foremost, the case of the database of the Museum Engelandvaarders shows that a sound data model based on standards can be set up in such a way as to preserve maximum flexibility for stakeholders to make new requests. Using semantic triples (or quads) as the functional model for data storage is an important part of that flexibility.

Furthermore, the case study is proof that lofty, elaborate standards thought out in academia or in government institutions can – with some careful scaling - be applied in the field, even in small institutions, and without government funding.

On a more personal level, the author was happy to notice that TEI, a standard he first encountered 25 years ago, can still offer useful insights.

References

[1] http://www.museumengelandvaarders.nl/museum-engelandvaarders-english/
[2] http://exist-db.org/exist/apps/homepage/index.html
[3] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html#NDPERS
[4] https://link.springer.com/content/pdf/10.1007%2F978-1-4419-0221-4_16.pdf
[5] http://www.cidoc-crm.org/Entity/e73-information-object/version-6.2
[6] http://www.cidoc-crm.org/Entity/e22-man-made-object/version-6.2
[7] https://pdfs.semanticscholar.org/1200/6b4563971515709cf4de82d9a2e05772fb28.pdf
[8] http://www.isni.org/
[9] http://authorities.loc.gov/
[10] http://www.geonames.org/
[11] https://www.w3.org/TR/NOTE-datetime (examples)
[12] https://www.w3.org/TR/rdf11-primer/#section-triple
[13] https://www.w3.org/TR/n-quads/
[14] http://www.loc.gov/standards/mets/
[15] http://www.loc.gov/standards/premis/

Author Biography

Martijn van der Kaaij is a founding partner at Heron Information Management LLP. As part of his master's degree in history, he studied the application of ICT to the arts and humanities, which developed into an enduring fascination.

Martijn has 20 years experience of delivering training on metadata, process management and workflows. He also provides consultancy on these subjects, and develops software for both quality control and management and dissemination of metadata.