Relational Modelling of Historical Data: A Technical Perspective

Christof Rauchenberger and Alexander Watzinger

Presented at the International Medieval Congress Leeds, July 5th 2016

Digitising Patterns of Power, IV (Session 803)

I - Introduction

In the summer of 2013, we were approached by Katharina Winckler and Stefan Eichert of the Austrian Academy of Sciences. They each had a collection of historical and archeological research data in separate databases. However, their databases were structured very differently, and they wanted us to find a way to merge them.

This initial requirement led to the idea for a web interface to facilitate the gathering of further data. They also wanted to make sure that their data would be usable in other projects.

Our cooperation has been going on for three years now. The most visible result so far has been the OpenAtlas web application, which we released as open source software in April 2016.

You might be under the impression that software development is a rather complicated activity. And you would be right, it can be quite tricky. However, it is important to bear in mind that every piece of software exists ultimately to solve a problem and one of the most difficult parts of software development is to understand the problem and find a solution.

One of our first tasks therefore was understanding the requirements of the scientists.

The two we will discuss in this paper are:

- Compatibility
- A web interface for the input of complex data

II - Compatibility

When we talk about compatibility, we mean that your data should be structured in a way that makes it easy to incorporate into other research, whether your own future projects or somebody else's. Before we delve into the details of our data model, however, let's examine some approaches to storing, processing and accessing your data.

The most basic type of data storage is pen and paper. This method has significant advantages: it is extremely mature (people have been writing stuff down for quite some time now, after all) and has practically no setup costs: just buy a notebook and start writing things down. Unfortunately, its drawbacks are also considerable: it can be quite difficult to find a specific piece of information and data gathered in notebooks is almost by definition unstructured. And of course everybody has their own way of writing things down, which will often be opaque to others.

So eventually you will hit the limit of what you can accomplish with of a pen and paper based system. At this point, you will probably switch to a spreadsheet. Spreadsheets have many of the same advantages of pen and paper. Spreadsheet software is quite mature. It is also practically ubiquitous on computers and increasingly on smartphones, so the barrier to starting is very low. For many research applications, the spreadsheet is ideal - it's easily searchable and makes it easy to structure your data. However, spreadsheets do not lend themselves well to complex relationships between data. And like the notebook, everybody's spreadsheet is different.

If these disadvantages hamper you in your research, you might consider using a database. This could be a personal database, such as Microsoft Access, FileMaker or SQLite, or a full-fledged database server like PostgreSQL , Oracle or MySQL .

These will require a certain investment in database expertise. That is, you will need to learn how to model your data in a way that makes it easy to search and interrelate, or engage the services of a database professional.

Once you have overcome this obstacle, you can reap the many benefits of databases. They are capable of representing data of considerable complexity, especially regarding relationships between entries. Database servers especially are capable of accommodating many users simultaneously and include a whole rafter of tools to ensure the integrity of your data. The only disadvantage that remains is one the database shares with the spreadsheet and the notebook: your data model is going to be uniquely suited to your research, and not necessarily anybody else's.

That is why the projects involved in OpenAtlas, such as *Digitising Patterns of Power* and *Mapping Medieval Conflicts*, insisted on compatibility as one of their main goals. Fortunately Stefan Eichert already had a solution in mind: the CIDOC CRM.

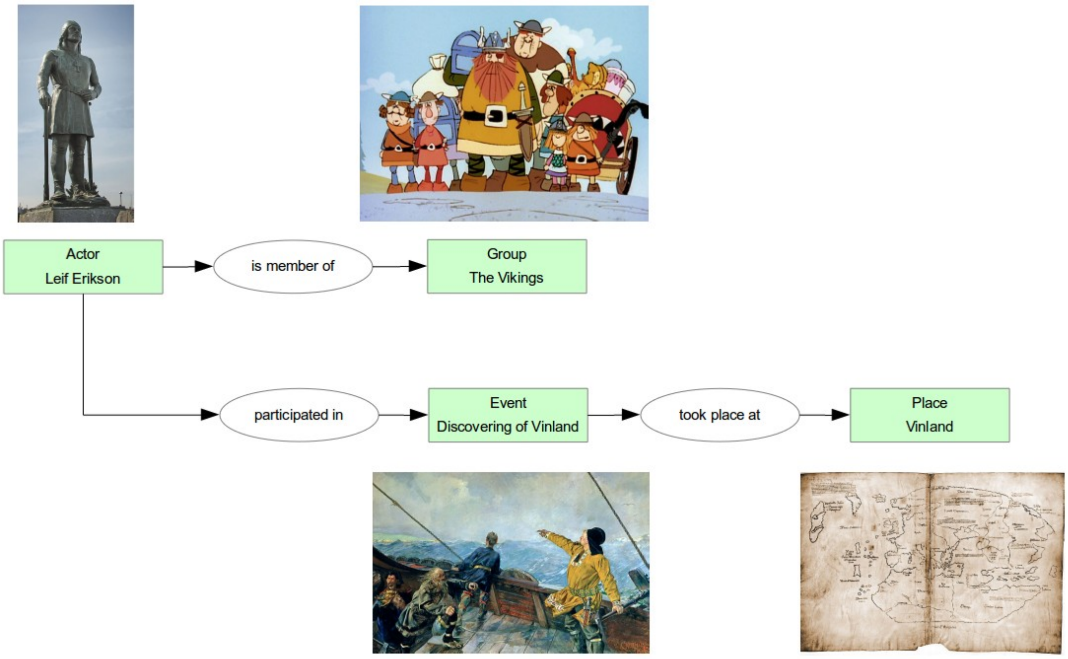
The CRM (Conceptual Reference Model) is a standardised way of structuring your data so it is accessible to every other data collection that also uses the CRM. It has been under constant development by the International Council of Museums since its inception in 1999, and was standardised as ISO 21127 in 2014. Its extensibility and compatibility stem from its object-oriented approach - every datum (such as dates, people, and countries) is considered an autonomous unit or Entity, and Entities can be connected via Properties.

III - A Web Interface for Data Entry

Developing OpenAtlas is a continuous process. From the beginning on we at *craws.net* worked in close cooperation with the members of projects using the software. There were online prototypes, many meetings and interesting discussions. All project members participated in the development, especially Stefan Eichert, who had already done a lot of the conceptual work.

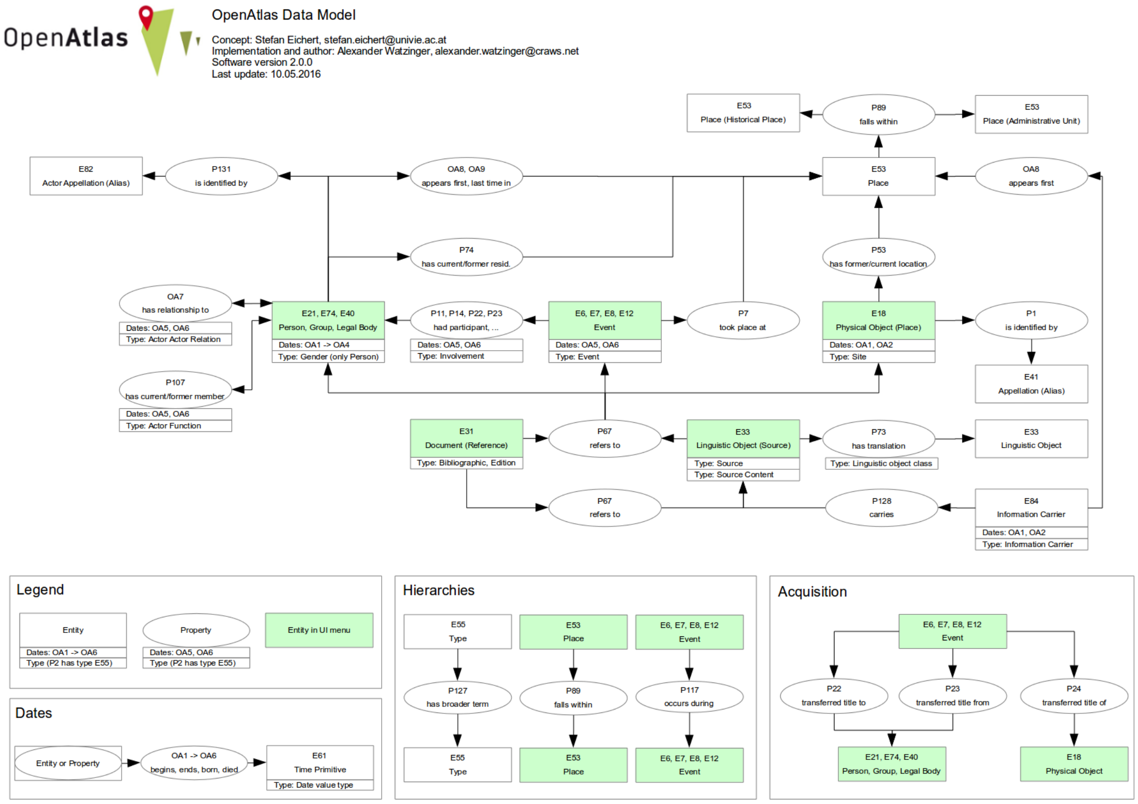
Guarantee CIDOC CRM conformant data integrity

One of our first tasks was to guarantee the data integrity. At the core the CRM describes entities which have classes like actor, event and place, and rules on how to link these classes via properties. As an example we will use *Leif Erikson*.



As you can see we have an actor (Leif Erikson) who is a member of a group (The Vikings). He also participated in an event (Discovery of Vinland) that happened at a place (Vinland).

On the next image you can see an already simplified version of our model to give you an impression of what we are dealing with.



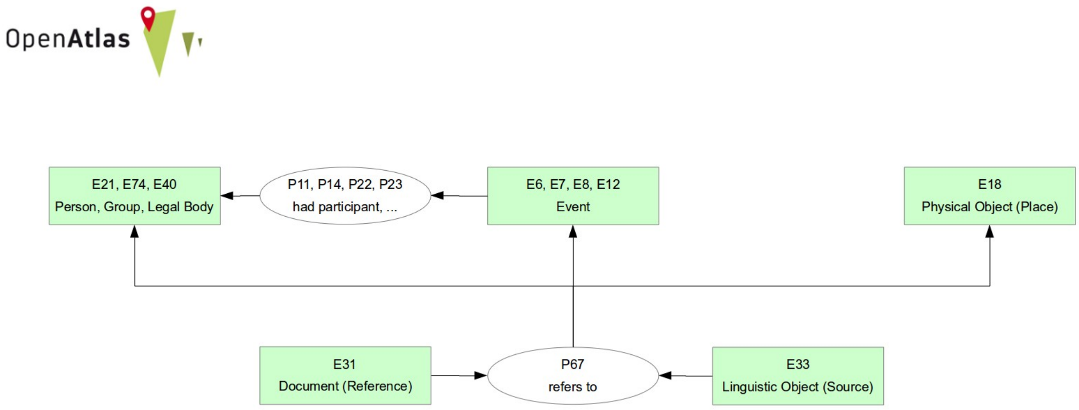
The CIDOC CRM specifies over hundred properties and rules how to use them so it can get quiet complicated to map all the information you are interested in. To guarantee data integrity we imported the whole specification into our system. It can be browsed in the web application and every link is automatically checked for validity.

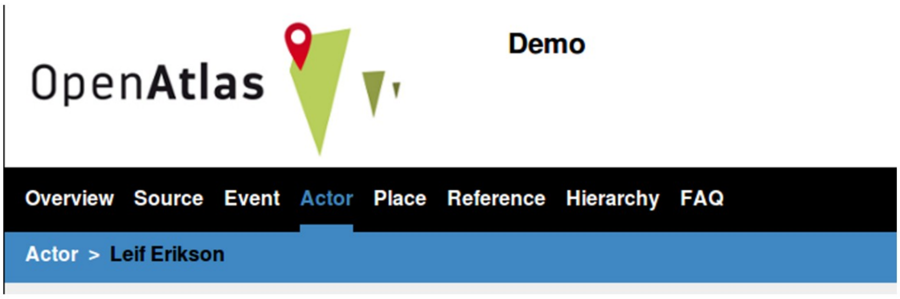
So now even if we make a design error the application will tell us immediately.

Ease of Use

The next step was to design a web interface where users don't have to be aware of the underlying model. We looked at typical use cases to define the entry points and took them as our main menu points.

As you can see on the next image they are: Source, Event, Actor, Place and Reference





From there you can create entries and add various connections. For example you can go to *Actor*, create or edit *Leif Erikson* and add the event *Discovery of Vinland*.
Or the other way round - you start at event, select *Discovery of Vinland* and add *Leif Erikson* from there.

Combined with a site wide search it is pretty easy to navigate and manage your data even without knowledge of the underlying model.

Flexibility for Different Research Topics

One important consideration was to keep the application flexible enough so it can be used in different projects with different research contexts.

For that purpose we made massive use of types which can be linked to Entities.

For example: for persons there is a predefined type *Gender* with the options *Female* and *Male*.

Types are hierarchical and can be extended by users so one could add the subtypes *Alpha*, *Beta* and so on to a gender, or even introduce new ones if your research requires it.

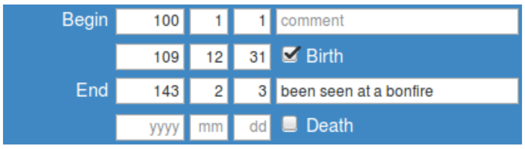
With the recent addition of dynamic types it is even possible to add new types, for example a type *Haircolor* for persons. They can even be used for multiple entities for example a type *Prominence* could be added to Actor, Place and Event.

Dealing with data fuzziness

Often the available data is incomplete or fuzzy. For example the name, lifetime and location of a person could be vague.

For dealing with names we implemented aliases so that different spellings of names can be recorded.

To manage uncertainty in time we are using up to 4 dates. A timespan for the beginning and a timespan for the ending.



Begin	100	1	1	comment
	109	12	31	☑ Birth
End	143	2	3	been seen at a bonfire
	yyyy	mm	dd	□ Death

Additionally there is an option for persons to choose if a date (or timespan) is the birth/death or the first/last appearance.

Right now, we are working on a feature in cooperation with *Digitising Patterns of Power* that will let us map uncertain locations. For example an old church could have a coordinate location, a shape if building plans are available or an area where it was believed to have existed. It can even be a combination of all of them.

These tools provide you with many options to record even fuzzy data, without loss of information.

Further Information

OpenAtlas	http://openatlas.eu
Demo	http://openatlas.craws.net
DPP	http://dpp.arz.oeaw.ac.at
MEDCON	http://oeaw.academia.edu/MappingMedievalConflict
CIDOC CRM	http://www.cidoc-crm.org