Bio CRM: A Data Model for Representing Biographical Information for Prosopography

Jouni Tuominen, Eero Hyvönen
Aalto University, Semantic Computing Research Group (SeCo)
Project homepage: http://seco.cs.aalto.fi/projects/biographies/

Goal: Create a systematic way to represent biographical information for prosopography, including basic data of people, personal relations, occupations, and events with participants in different roles.

Use Cases for the Data

Information retrieval. Find a set of people that share selected characteristics. For example: men who were born in England 1800–1850 and had certain occupations. This set constitutes a group of special research interest and can be analyzed further or compared with other groups. In this use case a SPARQL SELECT is typically used to create a tabular set of selected instances.

Network analysis. Find a network of people based on some criteria. This network can then be analyzed further by network analysis tools or visualizations. In this use case a SPARQL CONSTRUCT can be used for creating an RDF network based on the underlying data.

Knowledge discovery. Identify (automatically) a hidden group of people in the data that share some characteristics of interest that are not known beforehand [1]. In this use the result of a SPARQL SELECT or CONSTRUCT is analyzed further by specific algorithms or visualization tools.

Dynamic analysis. Analyze the structure and changing composition of a group and the changing roles of individuals or subgroups [1]. In this use the result of a SPARQL SELECT or CONSTRUCT is analyzed further by specific algorithms or visualization tools.

Design Principles

Bio CRM provides a semantic data model for harmonizing and interlinking heterogeneous data from different biographical data sources.

A natural choice for modeling life stories is the event-based approach where a person's life is seen as a sequence of spatiotemporal, possibly interlinked events from birth to death. A person may also be involved in prenatal and posthumous events. However, it is not necessarily clear what is an event. For example:

- Are occupations events, such as being a bishop or a president, because holding an office occurs in time and space with an agent involved?
- Are family relationships events, e.g., getting and being married?

The concept of, e.g., “bishop” would be useful in representing and querying biographical data, and there are indeed bishops around in the real world, but what does an instance of the class bishop mean? And how does it relate to the event of holding a bishop's office? If being a bishop can be represented in different ways, querying of data conforming to such a model becomes very difficult since the user cannot be sure in what alternative ways different bishops are actually represented.

To avoid such confusions, Bio CRM makes a clear distinction between the person's attributes, relations between people, and events in which people participate in different roles. These are characterized in the following way:

- Attributes are properties of a person that are assumed to characterize her independently of time and space, e.g., occupation.
- Relationships are established between people and are assumed to characterize the people independently of time and space, e.g., father-of.
- Roles express ways in which people participate in events, e.g., a baptismal candidate.
- Events take place in time and space and involve participants in different roles.

Also the attributes and relationships are modeled systematically using a role-based pattern.

The **core design principles** of the data model are:

- The model is a domain specific extension of CIDOC CRM, making it applicable to not only biographical data but to other CH data, too.
- The model makes a distinction between enduring unary roles of actors, their enduring binary relationships, and perduing events, where the participants can take different roles modeled as a role concept hierarchy.
- The model can be used as a basis for semantic data validation and enrichment by reasoning.
- The enriched data conforming to Bio CRM is targeted to be used by SPARQL queries in a flexible ways using a hierarchy of roles in which participants can be involved in events.

Alternative approaches for modeling the roles of participants in the events is discussed in the document:

https://docs.google.com/document/d/1XplgUBjCz16nYwhmLwxKV5av6dn3QHz5o8pMN6jTUbg

Bio CRM's role modeling is based on the Alternative 3: "instantiating roles".

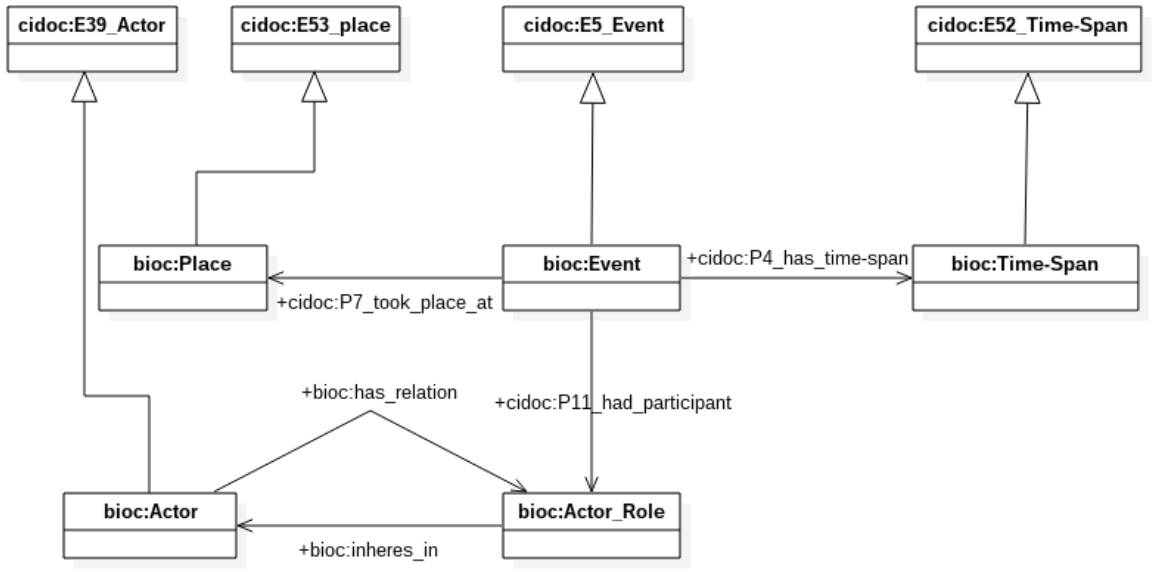
Bio CRM's representation of time is discussed in the document: https://docs.google.com/document/d/1wKYc3-cnKH7GIKCnNhlGhwfwFdLzPEI0ZEgug9sh5RI

Core Classes of the Model

The core classes of the model are organized into the following hierarchy:

Entity
- Actor
  - Person
  - Group
- Actor_Appellation
- Document
- Entity_Role
  - Actor_Role
  - Thing_Role
- Event
- Place
- Thing
- Time-Span

The relations between the core classes are presented in the diagram:



In the following, the core classes are described.

The namespace URI for the Bio CRM is *bioc*: http://ldf.fi/schema/bioc/

Technical schema documentation and the OWL ontology is available at the namespace URI: http://ldf.fi/schema/bioc/

Schema documentation in a spreadsheet format is available in the document: https://docs.google.com/spreadsheets/d/1suY8sjM6zLtxQ6ywAFVPmlBHpri0An_bOgGzC_0Ve4w

Person

A central focus in representing biographical data is to represent people and their networks. A person is represented as an instance of bioc:Person, a subclass of cidoc:E21_Person. This instance-of relationship is persistent and never changes during the life of the person.

In order to identify a person, the person is associated with core data:

1. Appellations, i.e., names and identifiers in other data repositories
2. Birth time and place
3. Death time and place

A person can also have other attributes, relationships, and participate in events:

4. Unary roles
    a. Gender
    b. Nationality
    c. Occupation

5. Relationships
    a. Family relations
    b. Social relations
    c. Group relations
6. Events

Appellations

A person may have lots of appellations, similarly as in ULAN [2]. Firstly, appellation may be of different kinds, e.g., Pseudonym, Nickname, etc. These are represented as subclasses of bioc:Actor_Appellation (which is a subclass of cidoc:E82_Actor_Appellation):

bioc:Actor_Appellation
    bioc:Official_Name
    bioc:Alternative_Name
        bioc:Nickname
        bioc:Pseudonym

cidoc:P131_is_identified_by is used to associate an actor with an appellation.

The literal form of a name in an appellation is given by property rdfs:label. A name may be transliterated differently in different languages, which can be represented by using language tags. A name can be valid at different times, e.g., due to marriage), which is represented using cidoc:P4_has_time-span. E.g.:

```
:jacqueline_kennedy_onassis a bioc:Person ;
  cidoc:P131_is_identified_by :jacqueline_kennedy_onassis_birth_appellation,
:jacqueline_kennedy_onassis_after_marriage_to_jfk_appellation .

:jacqueline_kennedy_onassis_birth_appellation a bioc:Official_Name ;
  rdfs:label ”Jacqueline Lee Bouvier” .

:jacqueline_kennedy_onassis_after_marriage_to_jfk_appellation a bioc:Official_Name ;
  rdfs:label ”Jacqueline Lee Kennedy” .
```

Birth and Death

Person’s birth and death are represented as a Birth/Death event, which can be qualified with time and place. Birth can also incorporate information about the mother and father. E.g.:

```
:john a bioc:Person .

:birth_of_john a cidoc:E67_Birth ;
  cidoc:P98_brought_into_life :john ;
```

```
  cidoc:P4_has_time-span :time1 ;
  cidoc:P7_took_place_at :place1 ;
  cidoc:P96_by_mother :johns_father ;
  cidoc:P97_from_father :johns_mother .

:death_of_john a cidoc:E69_Death ;
  cidoc:P100_was_death_of :john ;
  cidoc:P4_has_time-span :time2 ;
  cidoc:P7_took_place_at :place2 .
```

Unary Roles

Having a role is, say Teacher, may be something passing or something inherent characterizing a person as a whole in all times, even if it is possible also to specify when exactly the role was present (e.g., a professorship). Being a Teacher by education is different from saying that the person happened to participate in a particular teaching event, e.g., gave a lecture, in the role of a Teacher.

Genders, nationalities, and occupations of people are represented by relating a person to a unary role using the property bioc:bearer_of. While this expresses the gender, nationality, or occupation generally, it's also possible to qualify the roles in time and space by attaching a contextualizing event, e.g., the employment of a person. This is useful, as people have different roles during their life that typically perdue a shorter period of time and may have other qualifiers, too. For example, John Kennedy had the role of President in the US in 1961–1963. See further discussion on using events to contextualize roles in the section "Events".

The classes for representing people in unary roles for genders, nationalities, and occupations are organized in the following class hierarchy:

```
bioc:Actor_Role
        bioc:Unary_Role
                (Simple roles involving only one participant)
                bioc:Gender
                        e.g., Male, Female
                bioc:Nationality
                        ... (defined in individual prosopography datasets)
                        e.g., British, Finnish
                bioc:Occupation
                        ... (defined in individual prosopography datasets)
                        e.g., Philosopher, Merchant
```

Gender

The gender of a person is expressed using the bioc:has_gender property (subproperty of bioc:bearer_of). E.g.:

```
:john a bioc:Person ;
  bioc:has_gender [ a :Male ] .
```

Nationality

The nationality of a person is expressed using the bioc:has_nationality property (subproperty of bioc:bearer_of). E.g.:

```
:john a bioc:Person ;
  bioc:has_nationality [ a :British ] .
```

Occupations

The occupation of a person is expressed using the bioc:has_occupation property (subproperty of bioc:bearer_of). E.g.:

```
:john_kennedy a bioc:Person ;
  bioc:has_occupation [ a :President ] .
```

Relationships

The same role-based pattern is used for representing inherent relationships between people, such as family relations (mother, cousin, aunt, etc.) and social relations (studentOf, knows, etc.). Relationships are represented by relating an actor (a person or group) to another actor *in a role* by using one of the subproperties of the property bioc:has_relation. For example:

```
:john_kennedy a bioc:Person ;
  bioc:has_family_relation [ a :Spouse ; bioc:inheres_in :jacqueline_kennedy_onassis ] .
```

Similarly as unary roles, relationships can be qualified with temporal and spatial information by using an event to contextualize the role. A person may have been at some point a Spouse, a Lawyer in a company, and a President of a country, possibly several times at different occasions. For example, John Kennedy was Spouse of Jacqueline Kennedy in 1953–1963.

The classes for representing people in roles for personal relations are organized in the following class hierarchy:

```
bioc:Actor_Role
        bioc:Binary_Relationship_Role
                (Roles involved in binary relationships)
                bioc:Person_Relationship_Role
                        bioc:Family_Relationship_Role
```

... (defined in individual prosopography datasets)
e.g., Aunt, Husband
bioc:Social_Relationship_Role
... (defined in individual prosopography datasets)
e.g., Teacher, Follower
bioc:Intergroup_Relationship_Role
... (defined in individual prosopography datasets)
e.g., Subsidiary, Holding_Company
bioc:Group_Relationship_Role
... (defined in individual prosopography datasets)
e.g., Host_Organization, Member

The properties for expressing relations between actors are organized in the following property hierarchy:

bioc:has_relation
bioc:has_group_relation # between person and group
bioc:has_intergroup_relation # between groups
bioc:has_person_relation # between people
bioc:has_family_relation
bioc:has_social_relation

Events

The individual events of biographies are represented as subclasses of bioc:Event that is a subclass of cidoc:E5_Event inheriting its properties. From a semantic viewpoint, events are described especially in terms of

- the time of the event (cidoc:P4_has_time-span),
- place of the event (cidoc:P7_took_place_at),
- actors that participated in it (cidoc:P11_had_participant),
- other resources involved (cidoc:P12_occurred_in_the_presence_of).

Time and place properties refer directly to time spans and instances of places, respectively. The values for participating actors and other resources are instances of role classes. An actor role associates an actor with a role, making it possible for a person to participate in events in different roles that can also be qualified in terms of additional properties. Similarly as actors, physical objects and immaterial things can be involved in an event on specific roles. For example, a teaching event could be the following:

```
:t12 a :Teaching_Event ;
  cidoc:P11_had_participant [ a :Teacher ; bioc:inheres_in :person35 ] ;
  cidoc:P11_had_participant [ a :Student ; bioc:inheres_in :person87 ] ;
  cidoc:P4_has_time-span :time43 ;
  cidoc:P7_took_place_at :place78 .
```

Events can also be used for qualifying a unary (e.g., an occupation) or family relation further, i.e., in such cases an instance of bioc:Event has to be created. For example:

```
:jfk_presidency a :Hold_Office ;
  cidoc:P4_has_time-span :time_1961-1963 ;
  cidoc:P11_had_participant :jfk_in_role_president ;
  cidoc:P11_had_participant :usa_in_role_organization .

:jfk_marriage a :Marriage ;
  cidoc:P4_has_time-span :time_1953-1963 ;
  cidoc:P11_had_participant :john_kennedy_in_role_husband ;
  cidoc:P11_had_participant :jacqueline_kennedy_onassis_in_role_wife .
```

By using roles, it is possible to keep the number of properties smaller, because different properties for different roles are not not needed. Instead, different role classes are used. Such a model is simpler to query using SPARQL and provides the user with a set of useful and natural hierarchy of role concepts.updater

Event Hierarchy

The classes for representing events are organized in the following class hierarchy:

bioc:Event
- ... (defined in individual prosopography datasets)
- e.g., Confession, Baptism

For example, in a version of the EMLO data model following event classes are anticipated:

bioc:Event
- :Ecclesiastical_Event
- :Educational_Event
- :Political_Event
- :Professional_Event
- :Social_Status_Change

The class :Ecclesiastical_Event could be divided further into following subclasses with attached roles, such as:

- ;Baptism
  - **Roles:** :Officiant, :Baptismal_Candidate, :Godfather, :Godmother, :Religion
- :Confirmation
  - **Roles:** :Officiant, :Confirmation_Candidate, :Religion

Role Hierarchy

The classes for representing actors in roles participating in events are organized in a class hierarchy:

```
bioc:Entity_Role
        bioc:Actor_Role
                bioc:Event_Role
                        (Roles for actors in events)
                        ... (defined in individual prosopography datasets)
                        e.g., Officiant, Guest
        bioc:Thing_Role
                (Roles for other resources in events)
                ... (defined in individual prosopography datasets)
                e.g., Religion
```

Possible roles that can be attached to events are specified using the OWL restriction owl:AllValuesFrom on property cidoc:P11_had_participant. It is recommended that each event class, say Baptism, has a corresponding class of allowed roles, say Baptism_Actor_Role. Its subclasses are roles whose instances can be used in filling the roles:

```
:Baptism_Actor_Role rdfs:subClassOf bioc:Actor_Role .
:Officiant rdfs:subCassOf :Baptism_Actor_Role .
:Baptismal_Candidate rdfs:subCassOf :Baptism_Actor_Role .
:Godfather rdfs:subCassOf :Baptism_Actor_Role .
:Godmother rdfs:subCassOf :Baptism_Actor_Role .
:Religious_Thing rdfs:subClassOf bioc:Thing_Role .
:Religion rdfs:subCassOf Religious_Thing .
```

For example:

```
:Baptism a owl:Class ;
        rdfs:subClassOf [ a owl:Restriction ;
                                owl:onProperty :P11_had_participant ;
                                owl:allValuesFrom :Baptism_Actor_Role ] ;
        rdfs:subClassOf [ a owl:Restriction ;
                                owl:onProperty cidoc:P12_occurred_in_the_presence_of  ;
                                owl:allValuesFrom :Religion ] .
```

Here :Baptism_Actor_Role is a class whose subclasses consist of all actor roles related to a baptism event. It could also be possible to use union expressions (owl:unionOf) in owl:allValuesFrom constraints, if a named class is not needed. In this case the new class is potentially useful for querying, e.g., all actors related to baptism events.

In this way, the data annotator can be guided to use only the right roles, and the new role class can be used for finding resources in roles when querying. The role hierarchy facilitates sharing roles between events and modifying their role structure easily by just editing the role hierarchy. This is more flexible than, e.g., changing property names, if roles were represented using different properties.

In the same vain, roles for other things (bioc:Thing_Role) in events can be specified.

Note: the property bioc:inheres_in is used for representing both atemporal (unary roles and binary relationships without qualifiers) and temporal (qualified by using events) roles of people. This is an informed decision for the simplicity of the model. Contrasting approach has been chosen in Basic Formal Ontology (BFO) 2.0, where relations can be represented as continuant or occurrent, with separate relation types for them [3]. This approach has been criticized by one of the OBO Relations Ontology (RO) developers (the relations have been historically ceded from RO to BFO) for its complexity, causing logic and usability issues [4].

Full hierarchy of all the roles in the data model:

bioc:Entity_Role
- bioc:Actor_Role
  - bioc:Unary_Role
    - (Simple roles involving only one participant)
    - bioc:Gender
      - E.g., Male, Female
    - bioc:Nationality
      - ... (defined in individual prosopography datasets)
      - E.g., British, Finnish
    - bioc:Occupation
      - ... (defined in individual prosopography datasets)
      - E.g., Philosopher, Merchant
  - bioc:Binary_Relationship_Role
    - (Roles involved in binary relationships)
    - bioc:Person_Relationship_Role
      - bioc:Family_Relationship_Role
        - ... (defined in individual prosopography datasets)
        - e.g., Aunt, Husband
      - bioc:Social_Relationship_Role
        - ... (defined in individual prosopography datasets)
        - e.g., Teacher, Follower
    - bioc:Intergroup_Relationship_Role
      - ... (defined in individual prosopography datasets)
      - e.g., Subsidiary, Holding_Company
    - bioc:Group_Relationship_Role
      - ... (defined in individual prosopography datasets)
      - e.g., Host_Organization, Member
  - bioc:Event_Role

(Roles for actors in events)
... (defined in individual prosopography datasets)
e.g., Confession_Giver, Guest

bioc:Thing_Role
(Roles for other resources in events)
... (defined in individual prosopography datasets)
e.g., Religion

References

1. Robin Buning: Prosopographical Research Questions. Presentation at the COST RRL WG2 Workshop on Biographical Linked Data and Prosopography, Oxford, January 22, 2015. http://seco.cs.aalto.fi/events/2016/2016-01-22-oxford/slides/2016-01-22-Buning-Oxford-WG2.pdf
2. The J. Paul Getty Trust: Union List of Artist Names (ULAN). http://www.getty.edu/research/tools/vocabularies/ulan/about.html
3. Barry Smith et al.: Basic Formal Ontology 2.0 – Specification and User's Guide. June 26, 2015. https://github.com/BFO-ontology/BFO/raw/master/docs/bfo2-reference/BFO2-Reference.pdf
4. Christopher J. Mungall: A critique of temporalized relations. v1.3. April 25, 2013. https://github.com/cmungall/trel-crit/raw/master/trc.pdf