

CATCH Full Proposal
600.640.000.10N16

October 18, 2011

1 Project Information

- 1a) **Project Title:** Mining Social Structures from Genealogical Data
- 1b) **Project Acronym:** MISS
- 1c) **Principal Investigator:** dr. Karl Tuyls (Maastricht University)

2 Summary

The starting point of this research project is the large collection of historical documents maintained by the Brabant Historical Information Center (BHIC). A document can be anything ranging from scans of birth and death certificates, memories of succession, or tax declarations, to official photographs or family pictures. The current status of this collection is that the documents have been tagged by source and subject; for example, birth certificates have been tagged by the name of the child, father, and mother, as well as the place and date of birth, and the source of the information. Researchers can use keyword-based search to find relevant documents for their research (either a scan or a pointer to a physical location) based on these tags. This database, however, is not at all flawless; many names are duplicate, have several alternative spellings, or even contain mistakes. The data contains inconsistencies, and some information is missing. Furthermore, important semantic links such as the parent-child relation are only implicitly available, making simple tasks such as finding out if two given persons are related, very labor intensive. This project proposal therefore addresses the problem of how to derive identities of persons and social structures from large sets of genealogical data available as text and photographs with incomplete information. In order to do so we want to investigate and deploy a combination of techniques from data mining, machine learning and human computation. The project goals are (a) a semantically enriched and cleaned version of the current database of the BHIC; (b) the development of advanced search tools to support historical research; and (c) providing automatic tools for supporting large scale prosopographical research. The project will be carried out in close collaboration with BHIC.

3 Classification

- I. Analysis, formalization and methodology.

4 Composition of the Research Team

Name and title	Affiliation	Expertise	Role
Academic Partners			
dr. K. Tuyls	Maastricht U	Machine Learning, Distributed Systems	Project leader; co-promoter AiO
dr. T. Calders	TU/e	Data Mining and Databases	Site leader TU/e; co-promoter AiO
dr. ir Kurt Driessens	Maastricht U	Machine Learning, (statistical) relational learning	Researcher
dr. George Fletcher	TU/e	Semi-structured data; RDF, OWL, SPARQL	Researcher
Prof. dr. Gerhard Weiss	Maastricht U	Knowledge Systems, Multi-Agent Systems	Promotor PhD AiO
Prof. dr. Paul De Bra	TU/e	Hypermedia structures	Promotor PhD AiO
dr. Christoph Bartneck	Canterbury U.	Human-Technology Interaction	Advisor
Core members—Cultural Heritage			
Rien Wols	BHIC	Industrial advisor, Project-relevant data and technology	Site leader BHIC; Co-supervisor PhD student/postdoc
Jacques van Rensch	RHCL	Industrial advisor, Project-relevant data and technology	Advisor historical stud- ies
Core members—financed by NWO CATCH fund			
postdoc (NN)	TU/e, BHIC, and Maastricht U		Researcher
AiO (NN)	TU/e, BHIC, and Maastricht U		Researcher
programmer (NN)	BHIC		Implementation

The project is a close collaboration between the academic partners Maastricht University (UM) and the TU Eindhoven (TU/e) on the one hand, and the Brabants Historical Information Center (BHIC) on the other hand. The PhD student will follow a bi-diploma trajectory at UM and TU/e. Due to Dutch legislation the official promoters of the PhD student needs to be a full professor. prof. dr. De Bra and prof. dr. Weiss will fulfil this role. Next to the administrative requirement, their respective expertise in hypermedia structures and knowledge systems is of critical importance for the project. The postdoctoral researcher and the AiO will spend at least two days a week on-site at the BHIC. The programmer will most likely be located at the BHIC full time for practical reasons of data access. Relevant additional expertise for the project will be provided by the Regional Historical Center Limburg (RHCL) and dr. Christoph Bartneck of the Canterbury University, who will, more specifically, provide input to task 6.

5 Research School

The research school involved is SIKS (School voor Informatie- en Kennissystemen).

6 Description of the Proposed Research

This section describes a research project for one Ph.D. student, one Postdoc and a technical programmer and is organized as follows: We start by discussing the problem area of prosopography and present the overall objectives of this project. We illustrate the needs of the researchers in this area with an illustrative example. The specific research challenges associated with the objectives are listed subsequently, and finally, we elucidate how we intend to approach each of the challenges provided in our methodology.

6a Scientific aspects

Problem statement and objective

Prosopography [36, 18] is the field that studies the common characteristics of a specific group of persons in order to unravel previously unknown information about and connections within this group under study. Such a target group exhibits one or more commonalities such as for instance relatedness by family ties, being part of a city council, having studied at the same university etc. Prosopography is therefore a social science encompassing genealogy, onomastics, demography, and sociography and is mainly practiced by (teams of) historians, linguists, politicians etc. Prosopographical researchers proceed by bringing together all relevant biographical data of a group of persons and looking manually for patterns in this data. This is done by relating information about, and searching for connections between different persons, revealing new historical and social insights.

One example of a type of prosopographical study is the identification of small towns in North Brabant in which during a period of several decades, all public functions were held by a limited number of families. In some of these small towns the title of major went from father to son and stayed within the same family for many generations. The identification of such local dynasties characterizes the socio-political situation of a community and provides us with important insights about the influence and power of certain families. For example, recent studies revealed that the brewery and industrial family Verstraaten provided for over a century the major of the town Mill starting in the 19th century ¹ and in the same period the brewery family Godschalx was in the major's seat for over a century in the town Berlicum ². Remarkable is that in both cases it concerns a family of brewers, indicating the great importance of this economical activity in the early 19th century.

The prosopographical research method requires the collection of a large amount of data of a well-defined group in order to discover these patterns or new information. The humanities research groups from the Brabants Historisch Informatie Center (BHIC) and the Regionaal Historisch Centrum Limburg (RHCL) contain interdisciplinary teams of historians, linguists and lawyers that perform this type of prosopographical investigations. Their work consists of assembling large amounts of data on a target group and manually searching through these documents in quest for prosopographical insights. These studies hence require an extensive, time-consuming search over multiple documents. When searching for local dynasties, for example, news paper articles need to be scanned to find clues about candidate towns to research more in depth, town hall and hospital registers for birth and death records to establish and confirm the family relation among subsequent majors of a town, legal notifications and official pictures to find who was the major of a town at a specific time. Incomplete or noisy information complicates this search further; e.g., in the Berlicum study the researchers had to take into account many spelling variants of the surname "Godschalx."

Both institutions have available vast amounts of genealogical data in the form of text and photographs with incomplete information, which grows significantly each day. Researchers can query this database, e.g. for a name, which returns a set of documents that needs to be processed and searched for patterns manually. Consequently, it is a time consuming and error prone process that could greatly benefit

¹<http://www.bhic.nl/index.php?id=11025> (in Dutch)

²<http://www.bhic.nl/index.php?id=10790> (in Dutch)

from an interdisciplinary cooperation with computer scientists and mathematicians, experienced in machine learning, formal reasoning and data mining. Such collaboration would allow for improving the consistency and quality of the data, partially automating the search process, and looking for patterns in a more systematic way. The purpose of this research project is exactly such an interdisciplinary cooperation between humanities and exact sciences. More precisely, The project will bring together scientists in the fields of machine learning, data mining, hypermedia structures, distributed systems and user interface modelling, as well as experts in history, language and law, with the following objective:

Objective

To develop a novel inter-disciplinary methodology for semi-automatically deriving, maintaining and using social structures from text and photographs using a combination of machine learning, data mining and human computation techniques.

The Need for Automation

Characteristic for the data of the BHIC and RHCL is the existence of many different types of objects with heterogeneous links tying them together. For example, *people* live in a *location*, have a *family relation* with other people, have a *profession* in a *company* or *institution*, *photographs* display *people*, *buildings* at certain *locations* at a specific *time*. *Towns* have *mayors*; every *major* holds the position for a certain *time span*.

In this project we claim that navigational support for this data collection would equip a prosopographical researcher with unprecedented analytical capabilities, enabling large scale research that would not be possible without this support. Imagine, e.g., a study of local dynasties of *all* towns under 5000 inhabitants in North-Brabant, where any city council role is passed from one family member to another (possibly with a different surname). Such large-scale studies are hardly feasible with the current technological support. To illustrate the power of automation, we sketch a realistic scenario illustrating the navigational and exploratory support of the MISS system we plan to realize in the project.

Scenario 1: The power of navigational support. *A researcher wants to study the class of important people in the cities around Eindhoven. For this purpose, the researcher indicates a couple of high-level professions, such as judges and city council members. She selects all people with a professions of this list in the period 1800-2000 in the Eindhoven area. Because she is uncertain whether she did not miss any important professions, she lists for all the people in that group and their first-degree family, the frequently recurring professions. From the acquired list she notices that also brewers should be added to the list of influential professions. After adding brewer, again all people with these professions are selected. In theory she could now repeat this step and “bootstrap” to get even more professions, yet she decides that for the purpose of her research the list is complete enough.*

Scenario 2: The power of exploratory support. *The second step of her research consists of determining common features among the important people. For this purpose, she uses the re-description engine of the MISS system. This engine finds alternative descriptions of a selected group in terms of characteristics selected by the researcher. The following interesting relationship is revealed: an accurate alternative description of the set of important people is: male and [(salary in top-5%) or (family Verstraaten) or (family Godschalx)]. This result triggers the researcher to take a closer look at the family Godschalx. She would like to know if this family is present in all high-level professions, or if they have a certain specialization. For this purpose, the discrimination module of MISS can be used: male Godschalx family members are selected and are compared to the group of all people having a high-level job. This step reveals that the distinguishing factor is that Godschalx family members are especially present in city-council functions.*

With the current technological support, these scenarios would take a significant amount of time from many researchers and moreover it is likely they would miss important patterns.

Research Questions

To meet the project objective, we identify the following research questions.

- **RQ1:** How can we automatically identify objects such as persons, locations, professions in the database and disambiguate duplicate entries? How to construct inference models and algorithms 'on-the-fly' using probabilistic machine learning techniques and reasoning systems, that can derive new knowledge about identities and social structures when new information about the context of a photograph, persons on such a photograph or relations between persons becomes dynamically available?

Some inconsistencies cannot be resolved by computationally techniques only; the following question relates to taking expert knowledge into account:

- **RQ2:** How to efficiently acquire new information from humans, i.e. relatives or acquaintances, that can be taken into account and provide crucial additional information for the machine learning and data mining loop? How to build appealing, efficient and un-intrusive user and interaction interfaces, allowing for acquiring new information by deploying smart human computation techniques?

Answering the first two research questions will lead to a semantically enriched and cleaned database. In order to support searching this database and enable the use of the semantic links, the following question needs to be answered next:

- **RQ3:** How can we provide navigational support for the semantically enriched database? What is a convenient way to present results to the researcher and to query the database, using, e.g., techniques such as faceted browsing? How can we support the offered operations at a technical level?

The final goal of the project is offering automatic tools for prosopographical research:

- **RQ4:** How can we efficiently discover patterns in large amounts of historical information, including text and photographs, providing insights in the context and relations expressed by the persons present in the data?

Approach

- **Method for RQ1** Question 1 deals with the automatic acquisition of a semantically enriched database. The first task will be a literature study:

Task 1

Both the Ph.d. student and postdoc perform an extensive literature study with respect to relevant scientific domains, including: common sense reasoning, inductive reasoning and probabilistic (relational) models. The technical programmer will get familiar with the ICT equipment and Memorix system used by the historical centers.

One of the corner stones of our approach towards the creation of a semantically enriched database will be computerized common sense reasoning [22]. When disambiguating names while identifying which records relate to the same person, common sense reasoning is essential. Before going into detail on the technical aspects of the computerization, we first illustrate a typical "common sense" deduction as it may be made by a researcher. Then we will explain how inductive and deductive techniques as developed in deductive database theory [31] and inductive logic programming [21, 11] will be applied to automate this process.

Table 1: All records returned by a keyword search for surname “Sitter”, early 19th century

Type	Surname	First name/Patronym	Role	Place	Date
death	Sitter, de	Gerhard Reincke	deceased	Beers	06-01-1824
	Sitter, de	Gerhard Cornelius	father of the deceased	Beers	06-01-1824
succ.	Sitter, de	Gerhard Reincke	deceased	Beers	06-01-1824
birth	Sitter, de	Gerhard	child	Beers	06-01-1824
	Sitter, de	Gerhard Cornelis Reincke	father	Beers	06-01-1824
birth	Sitter, de	Alida Philippina Johanna	child	Beers	31-01-1825
	Sitter, de	Gerhard Cornelis Reincke	father	Beers	31-01-1825
death	Sitter, de	Gerard Cornelius Reincke	relation of the deceased	Beers	26-02-1825

Example: Common sense reasoning

Consider the records in Table 1 representing the search result for the name “Sitter” in the early 19th century over several databases, including birth and death certificates, baptism records, successions, legal status (e.g., weddings) in 52 villages and cities of North-Brabant³. The records in the table have been organized according to the originating documents; two birth certificates, two death certificates and one memory of succession. Note that apparently on 6/1/1824 both a child was born (Gerhard de Sitter) and a person died (Gerhard Reincke de Sitter). When we inspect both the death certificate and the birth certificate, we see that the father of the child, resp. deceased is different: Gerhard Cornelis Reincke for the former and Gerhard Cornelius for the latter. Given the small variation in the names, it is highly likely that the middle name “Reincke” was assigned incorrectly to the deceased, and that actually the death record represents a child that died at birth; a tragic event not uncommon in those days. The fact that in both records the mother is Johanna Louisa Frederika Frans confirms our hypothesis, as well as the absence of a baptism record. We also see that Gerhard Cornelis Reincke de Sitter becomes father again in 1825; hence it was not him who died on 6/1/1824. To further confirm that the middle name Reincke belongs to the father, we try to find the wedding record of de Sitter - Frans. A text-search does not find this record. Therefore, we search for all records in the early 19th century of Johanna Louisa Frederika Frans. This returns the two birth certificates of Table 1, but without a marriage record. Therefore, we display all records connected to last name Frans in the period 1820 till 1824. In this way we find the wedding record of Johanna Louise Frederika Frans with Gerhard Cornelis Reincke de Sitter—but the surname of the bridegroom is registered as “Reincke de Sitter.”

Clearly, a system dealing with the disambiguation of person and place names in birth and death certificates, will have to rely on an extensive set of rules that describes the background knowledge underlying the common sense reasoning. In the database community, there already exists a massive literature on deductive databases [31]. A well-known example of a language for expressing such information in the context of relational databases is Datalog. Recently we see an explosive interest in Datalog as a declarative language that can be evaluated efficiently [37, 17]. Similar languages exist for less structured types of information; e.g., RDF [19] or OWL [24]. In the Datalog/OWL approach, rules describe family relations and express consistency constraints; e.g., the following hypothetical rules describe the concept *Father* and the rule that a person needs to have the same surname as her father:

```

father(Father,Child) :- parent(Father,Child), male(Father).
surname(Child,Name) :- surname(Father,Name), father(Father,Child).
    
```

Enforcing such rule on the database would have resolved the appearance of the surname “Reincke de Sitter” in the example given above. Also more complex information could be encoded in such rules; for example that a father and mother on the birth certificate should effectively have a matching wedding certificate:

³These records have been found through the online frontend of the BHIC archive available at <http://www.bhic.nl/index.php?id=11454>

married(Father,Mother) :- birthcertificate(Bcert,Child), registeredfather(Bcert,Father),
registeredmother(Bcert,Mother).

Task 2

In a first phase the Datalog approach will be implemented on top of the current relational system. The postdoc and domain expert will be involved in setting up the rule sets, which will be implemented by the programmer.

Unfortunately, a Datalog or OWL-based approach requires that all rules are given beforehand by a user and assumes consistent and complete input data. Obviously the second rule would not always hold as people may remarry, or records may be missing. Capturing all such exceptions will be tedious and is not realistic for the complexity of the proposed project. Therefore, only in an initial phase we will rely on rules to be extracted from frequent interactions with the domain experts. In a later phase, we will use a combination of the deductive approach with induction. In Inductive Logic Programming [21, 11], the goal is to find logical sentences describing common patterns in a given dataset. Such an inductive approach could detect, e.g., that children often receive the names of their godfather and mother as second and third name, which may be very helpful in disambiguating names.

Task 3

In parallel with the Datalog task, the PhD student works on an ILP approach to automatically extract relevant rules from the database.

Nevertheless, even the inductive approach lacks one important feature: all rules are inherently probabilistic and should be used in that way; i.e., the rules will give evidence of why one way to disambiguate the data should be preferred over another, but never prove it in a mathematical sense. Sometimes different rules may lead to different conclusions. We plan to combine these different conclusions using recent probabilistic versions of the ILP paradigm [10].

Task 4

Together with the postdoc, the PhD student extends the ILP approach to make the rule evaluation probabilistic. Depending on the literature study and experience it may be decided that another probabilistic model class is more appropriate; e.g., Probabilistic Relational Models (PRMs) [13]. The result of this task is a framework to incorporate deductions from probabilistic rules that have been acquired inductively. This approach will be implemented by the programmer.

• **Method for RQ2** As a result of the probabilistic ILP approach different disambiguations will be suggested, each associated with a probability. In the cases where there are many equally likely options, we will consult the user to break the ties. For determining the exact trade-off between consulting the user not too often and improving information quality, we will use ideas from active learning [9, 14]; the active learning paradigm emerges from semi-supervised learning and assumes that only partially labeled data has been given. In order to learn concepts the system is allowed to ask a limited number of labeling requests to the user. Selecting only few examples whose labeling would lead to the highest gain in information is a non-trivial task which is central in this research area.

Task 5

The postdoc develops a method to select the most uncertain instances to be presented to the user. An attractive user interface will be developed and implemented by the programmer.

Furthermore, in this step we will also research how to take the human in the loop from a user interaction perspective, by deploying human computation techniques. For example, when we target a specific family, older family members or acquaintances can help in identifying persons on photographs of a number of generations ago. Turnbull et al. showed that test subjects quickly tire of lengthy surveys, resulting in inaccurate annotations. Therefore the student will delve deeper into human computation methods as introduced by Ahn et al. [3, 35]:

Task 6

The student investigates, supported by advisory project member dr. Bartneck, how to include humans in the loop by turning the gathering of additional information from humans into a game, in combination with the active learning component. Starting point is Peekaboom, a highly successful game that helps a learning algorithm through the support of human players [4].

MILESTONE 1: THE SEMANTICALLY ENRICHED AND CLEANED DATABASE HAS BEEN CREATED

• **Method for RQ3** The third research question concerns the development of the navigational support on top of the semantically enriched database. Looking at the first scenario, the following types of operations need to be supported:

- Manage lists persons, professions, locations, etc.
- Follow semantic links; e.g., for all people give their profession;
- Filter objects based on their links; e.g., select people with influential professions during a certain time-frame;
- Define meaningful notions for the user that may depend on the existence of several subsequent links, such as “family in the first degree”;
- Compute simple statistics for lists of objects such as averages, sums. E.g., count how many people in another given list practiced a given profession in 1850.

These operations are typical operations for semi-structured data models such as the eXtensible Markup Language (XML) [2] with its query languages XPath [6] and XQuery [8] or the Resource Description Format (RDF) [19] or the Web Ontology Language (OWL) [24] with query language SPARQL [28] or extensions [26, 5]. XML is mainly intended for the storage and management of hierarchically structured data, whereas RDF is intended for more loosely organized graph-structured data, based upon the idea of making statements about so-called “resources” in the form of subject-predicate-object expressions. Given the structure and aim of our database, RDF or one of its extensions seems a more natural choice [34]. The first challenge we face will be exactly this choice and the construction of a suitable database architecture. The scalability of the system will be a major concern; the semi-structured query languages are at the current time not performing as good as their relational counterparts. In this respect an alternative choice may be to support an RDF or OWL-like format indirectly on top of a relational system, and to implement a limited set of specialized operations instead of the full SPARQL query standard. The second challenge relates to the temporal aspect [29, 15] in the data which is not naturally supported by the RDF model. Therefore, as a design choice, it needs to be decided whether this aspect will be encoded as a separate property associated with every node in the RDF-graph, or if it will be incorporated by extending the model itself. [34]

Task 7

Design a suitable data storage and querying format for the semantically enriched data. The architecture should allow for navigational support and the temporal aspect. At the same time be scalable to large amounts of data. Implement the data model.

XML, RDF, and OWL are mainly intended as computer-interpretable descriptions of data objects, their relations and meta-data. As such, the data and the answers to navigational queries cannot be presented directly to the users, but must be transformed in a more human-friendly format. This format should be such that the structure and the relations between the objects is preserved as much as possible in the presentation format. Ideally the navigational operations map naturally to the presentation format; e.g., a graph-based model where nodes can be expanded by clicking on them.

Task 8

Develop an appropriate way to present the XML/RDF/OWL data and the answers to the queries; e.g., building upon the Tabulator browser for RDF data developed at MIT [7]. Implement the presentation layer.

MILESTONE 2: NAVIGATIONAL SUPPORT IS COMPLETE; THE SYSTEM CAN BE DEPLOYED BY RESEARCHERS AT BHIC

Once the second milestone is realized, the system is ready to be used by the prosopographical researchers. This system is already capable of supporting all operations sketched in Scenario 1 and represents already a huge benefit.

Task 9

Illustrate the usefulness of the system with navigational support by using it in a large-scale prosopographical research project.

• **Method for RQ4** With the semantically enriched database and the navigational support it is possible to complete the system with mining and statistical learning tools for discovering patterns and new insights on target groups. The main difference with RQ3 is that in RQ3 the emphasis was on user-driven exploration of the database; the user reveals interesting patterns by browsing through the database and asking concrete questions to the system. For RQ4, however, we aim at automatic support for the exploration itself; the system itself will search for interesting patterns in the data. The difference between these two paradigms was also illustrated in Scenario 1, representing a typical navigational approach, versus Scenario 2 in which the exploratory support was illustrated. The first task here will be the implementation of pattern mining [16] on the semantically enriched database. The goal of the pattern mining module is to automatically discover significant and surprising patterns in a subgroup of objects selected by the user.

Task 10

Extend existing pattern mining techniques to the semantically enriched database.

Another important task is the identification of re-descriptions [39, 30, 12]. Instead of searching for regularities and patterns in a group of objects such as in the previous task, the goal now is to find alternative descriptions of a given group of objects; e.g., the system could find that a group of persons which lived in the same neighborhood could alternatively be described by their income class.

Task 11

Develop techniques for re-description-mining for the semantically enriched database.

In the previous two tasks, the groups for which we wanted to discover patterns, or that needed to be re-described, were given externally by the user. In subgroup discovery [20, 38], this is not the case. Given some target attributes, subgroup discovery aims at the detection of groups of objects that behave significantly different w.r.t. the target attributes. A discovered subgroup may present an interesting subclass in the community.

Task 12

Develop subgroup techniques for the semantically enriched database.

These three tasks will be far from trivial, as most existing data mining and machine learning techniques for pattern mining, re-description discovery and subgroup discovery assume data to be given in only

one table. Some extensions to the multi-relational setting have been proposed, but most of these extensions assume that the connections between the tables are well defined and simple; e.g., recursive structures, such as parent-child relations that can be arbitrarily deeply nested are usually not supported.

MILESTONE 3: EXPLORATORY SUPPORT IS COMPLETED

Task 13

Writing of the Ph.d. thesis by student. Finalization and implementation of complete software package, and field study at historical centers by programmer.

Related Work

In this section we concisely describe related projects and research that are of relevance to MISS.

Related research can be found in the CommonSense project [22, 23]. The main difference is that the focus in CommonSense is on methods generating common sense knowledge, while in this project the purpose is to use common sense knowledge, provided by humans, in combination with machine learning and data mining technology. [22, 23, 25, 33].

The 6th framework project CLASS: Cognitive-Leven Annotation using Latent Statistical Structure aimed at the development of machine learning techniques that allow discovery of, amongst others, people in images, video and associated text. The resulting probabilistic models for the alignment of names and faces in news texts and their accompanying images can be used as one of the starting points in this project. [27]

The NSF and Microsoft Research funded project called 4D Cities on the spatio-temporal reconstruction from images offers insights on automatic inference of a temporal order of images. While in this project the images focus on buildings and other city landmarks, similar alignment techniques could be applicable for images related to humans as well. [1]

Entity resolution deals with the problem that databases contain imprecise references to entities from the real world. For example, multiple people can go by the same name and there may be different names which refer to the same person as well. The goal of entity resolution is to cluster the database references according to their associated real world entities. [32]

6b Multidisciplinary cooperation

The project will bring together scientists in the fields of machine learning, data mining, hypermedia structures, distributed systems and user interface modelling, as well as experts in history, language and law. The project team comprises some of Europes leading experts in the areas essential for the topic of this proposal. An important condition for the success of the project is fulfilled, i.e. the Brabants Historisch Informatie Center (BHIC) and the Limburg Regionaal Informatie Centrum will not only provide the necessary datasets, but will also actively participate in the project by providing the necessary expertise from the humanities point of view with respect to the type and form of data, the current usage of the historical information systems, and the urgent needs for new technology to mine for identification and genealogy trees. The postdoc, Ph.d. student and programmer will be working physically at the universities as well as at both historical centers.

6c Relevance

The results of the project proposal will allow the humanities experts of the historical centers to obtain more specific information from their huge amount of datasets. More precisely it will influence and benefit prosopographical research (prosopography researches the common characteristics of a historical group) that is being conducted at the historical centers. It will allow for more advanced applications such as e.g. the determination of the origins and common characteristics of governors over a period of time in a particular region or city. This project is not only highly relevant to the historical centers participating in this project proposal but is also of importance to museums and genealogical organizations that can rely on and build applications based on the research results and software developed in the project. The research proposed is also relevant to other fields, e.g. medical research in genetic disturbances. Medical doctors already consult databases of BHIC to find genealogical information about their patients. The research developed in this project can facilitate and make this type of applications more accurate.

6d Research utilization

The proposed project can rely on the participation of the project team of BHIC and RHCL that contains linguists, historians, archivists and lawyers. This team will closely cooperate with the scientific team and support them with their knowledge and expertise in the area of genealogy, language and history. This unique combination of humanities and exact sciences will provide a fruitful climate for progress as well in the prosopographical field as in the machine learning, data mining and human computation areas. Both institutes will make available their archives, study halls, and knowledge.

Next to the technological innovations w.r.t. the development of new techniques and algorithms, one of the results of the project will be a fully operational database system to support the prosopographical research at the BHIC and RHCL institutes. We expect that due to the navigational support the amount of time spent by the historians on tracking the right information will reduce significantly. Furthermore, the exploratory support will make it possible to conduct statistically founded prosopographical researches at a much larger scale than is the case nowadays.

The MISS project is also expected to have a large impact beyond the historical researches; with the ever growing amount of graph-structured data, more and more databases suffer from the same limitations as the data collected at BHIC and RHCL: data is incomplete and inconsistent, contains heterogeneous sets of objects with multiple links between them. Very concretely, we see a clear application perspective for organizing and navigation through scientific literature. Consider, e.g., the *bio-informatics* research field, where it became literally impossible to even keep pace with the tremendous amount of papers being published. Despite initiatives such as UniProt⁴ that stores information about proteins harvested from literature, it remains difficult or even impossible to get a clear overview of the available information and to browse through it. The settings here are quite similar as in the prosopographical context: there is a notion of a document; i.e., a research paper, and different types of objects described in these documents; i.e., proteins, (metabolic) pathways, experiments, interactions, etc. and relations between them; e.g., a paper may describe one protein interacting with another, or being part of a certain metabolic pathway. Similar techniques as developed in the MISS project could be applied here.

7 Description of the proposed plan of work

In Figure 1, the different components of the project have been visualized. At the top we see in light blue the currently existing system containing tagged documents; e.g., birth or death certificates. The user can query the tags with keywords only.

⁴<http://www.uniprot.org/>

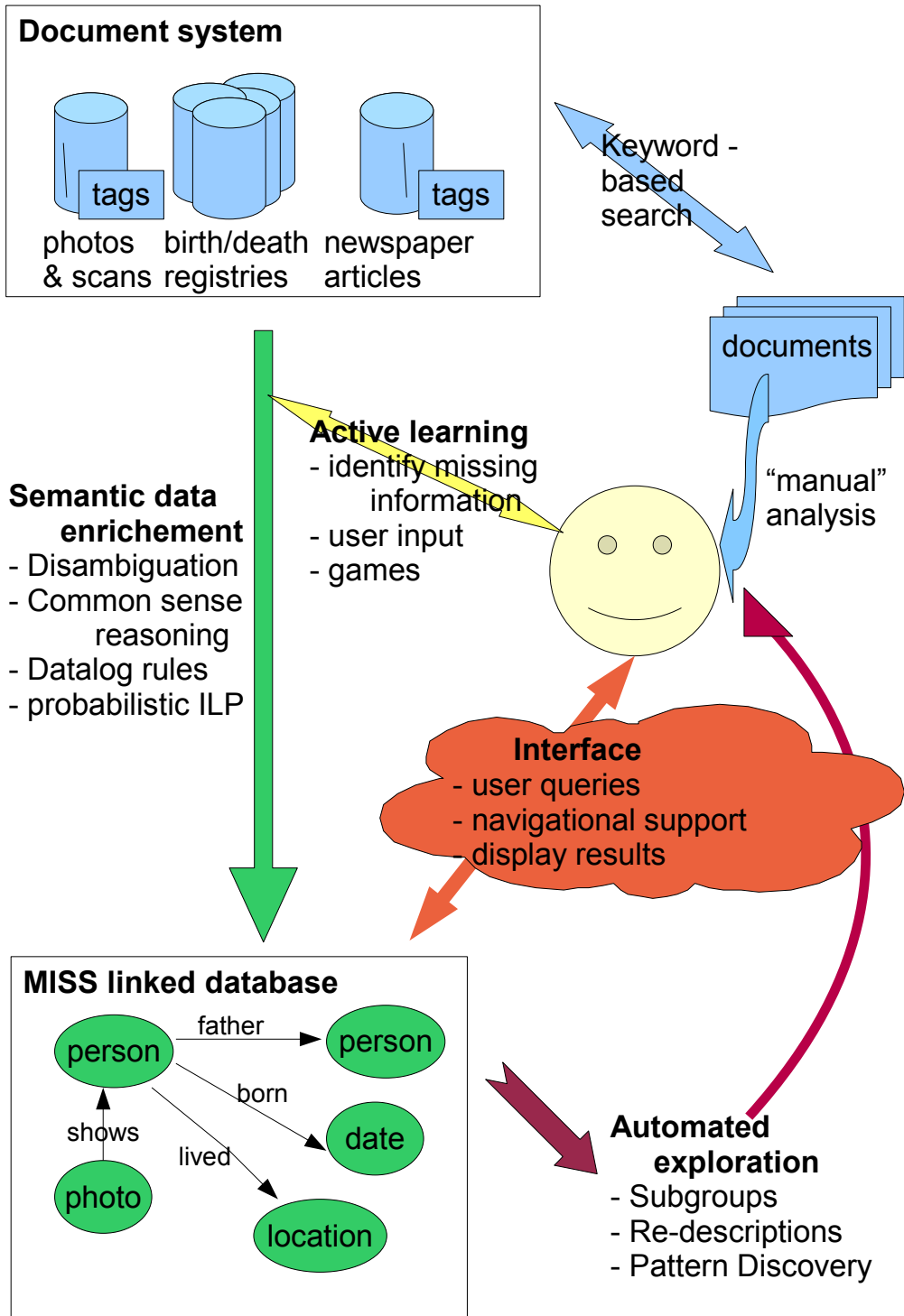


Figure 1: Graphical representation of the project approach

Workpackage WP1: Automatic Tools for Semantic Enrichment (dark green) Transform the data in the current system to the MISS database (at the bottom, green). WP1 consists of the development and implementation of all fully automatic tools based on machine learning and data mining techniques. It corresponds to research question RQ1 and encompasses tasks T1 till T4.

Workpackage WP2: User Interaction for Semantic Enrichment (yellow) For those cases where the computer is not able to resolve a conflict or cannot decide between two or more potential options, in the active learning setting, the user can be asked for input. The handling of user input in the semantic data enrichment is the content of WP2. It corresponds to research question RQ2 and encompasses tasks T5 and T6.

Milestone M1: MISS Database created Completion of WP1 and WP2 leads to Milestone M1.

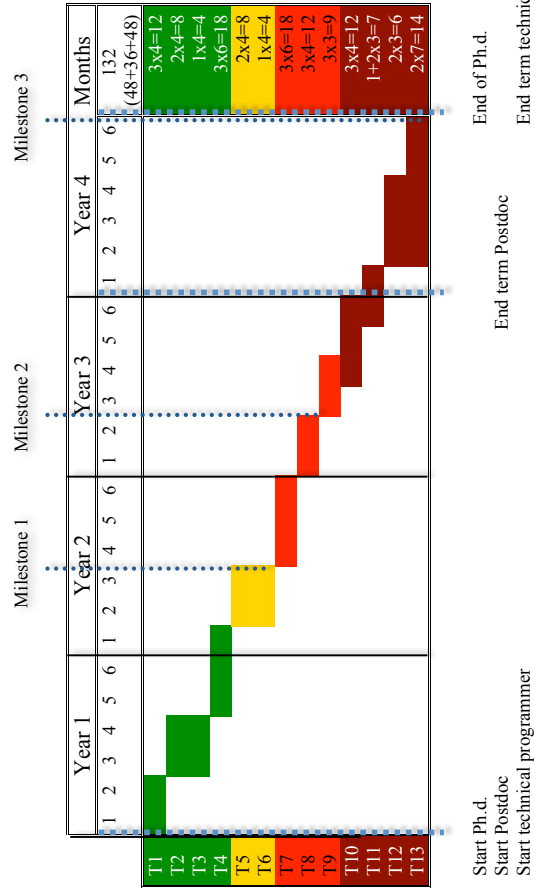
Workpackage WP3: Navigational Support (light red) Corresponds to RQ3; the interface for navigational support. This interface should allow for efficiently and transparently navigating along family trees, from persons to locations and back, to professions, documents, etc. It encompasses tasks T7 and T8 and leads to:

Milestone M2: Navigational Support. After reaching M2, the system can already be used to support prosopographical research. Task T9 will be carried out to support this claim.

Workpackage WP4: Exploratory support (dark red) Connected to RQ4; the development of inductive methods to automatically support prosopographical operations, such as group re-descriptions, pattern recognition (subgroup discovery), and discriminative analysis. This package poses several challenges from a data mining and machine learning perspective, since most knowledge discovery techniques have been developed either in the context of fully structured relational information structures, or homogeneous graph structures with a limited number of different labels. It encompasses tasks T10 till T13.

Milestone M3: Exploratory Support

To carry out the described work, we adhere to the project planning shown in the following gant chart. This chart contains the tasks outlined in the methodology and shows when they should be carried out and how much time is foreseen for each of them.



Student: T1 (4m), T2 (0m), T3 (4m), T4 (6m), T5 (0m), T6 (4m), T7 (6m), T8 (4m), T9 (3m), T10 (4m), T11 (3m), T12 (3m), T13 (7m), 48m
 Postdoc: T1 (4m), T2 (4m), T3 (0m), T4 (6m), T5 (4m), T6 (0m), T7 (6m), T8(4m), T9 (3m), T10 (4m), T11 (1m), T12 (0m), T13 (0m): 36m
 Programmer: T1(4m), T2 (4m), T3 (0m), T4 (6m), T5 (4m), T6 (0m), T7 (6m), T8(4m), T9 (3m), T10 (4m), T11 (3m), T12 (3m), T13 (7m), 48m

8 Literature

References

- [1] 4d-cities project. <http://www.cc.gatech.edu/4d-cities/dhtml/index.html>. Technical report, 2007.
- [2] S. Abiteboul, P. Buneman, and D. Suciu. *Data on the Web*. Morgan Kaufmann, 1999.
- [3] L. v. Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, 2008.
- [4] L. v. Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems, Montreal, Quebec*, 2006.
- [5] Faisal Alkhateeb, Jean-François Baget, and Jérôme Euzenat. Extending sparql with regular expression patterns (for querying rdf). *J. Web Sem.*, 7(2):57–73, 2009.
- [6] Anders Berglund, Scott Boag, Don Chamberlin, Mary F. Fernandez, Michael Kay, Jonathan Robie, and Jerome Simeon. Xml path language (xpath) 2.0 (second edition). Technical report, W3C Recommendation, December 2010.
- [7] T. Berners-Lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets. Tabulator: Exploring and analyzing linked data on the semantic web. In *Proceedings of the 3rd International Semantic Web User Interaction Workshop*, volume 2006, 2006.
- [8] Scott Boag, Don Chamberlin, Mary F. Fernandez, Daniela Florescu, Jonathan Robie, and Jerome Simeon. Xquery 1.0: An xml query language (second edition). Technical report, W3C Recommendation, December 2010.
- [9] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.
- [10] L. De Raedt and K. Kersting. Probabilistic Logic Learning. *ACM-SIGKDD Explorations: Special issue on Multi-Relational Data Mining*, 5(1):31–48, 2003.
- [11] S. Džeroski and N. Lavrač. *Relational data mining*. Springer Verlag, 2001.
- [12] A. Gallo, P. Miettinen, and H. Mannila. Finding subgroups having several descriptions: Algorithms for redescription mining. In *SIAM Data Mining Conf.(SDM)*, pages 334–345, 2008.
- [13] Lise Getoor and Ben Taskar. *Introduction to statistical relational learning*. MIT Press, 2007.
- [14] Z. Ghahramani, DA Cohn, and MI Jordan. Active learning with statistical models. *J. of Artificial Intelligence Research*, 4:129–145, 1996.
- [15] C. Gutierrez, C.A. Hurtado, and A. Vaisman. Introducing time into RDF. *IEEE Transactions on Knowledge and Data Engineering*, pages 207–218, 2007.
- [16] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007.
- [17] J.M. Hellerstein. Datalog redux: experience and conjecture. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems of data*, pages 1–2. ACM, 2010.
- [18] K Keats-Rohan. *Prosopography Approaches and Applications: A Handbook*. Oxford : Prosopographica et Genealogica, 2007.
- [19] O. Lassila, R.R. Swick, et al. Resource description framework (rdf) model and syntax specification. Technical report, World Wide Web Consortium W3C, 1999.

- [20] N. Lavrač, B. Cestnik, D. Gamberger, and P. Flach. Decision support through subgroup discovery: Three case studies and the lessons learned. *Machine Learning*, 57(1):115–143, 2004.
- [21] N. Lavrac and S. Dzeroski. *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood, New York, 1994.
- [22] Henry Lieberman, Hugo Liu, Push Singh, and Barbara Barry. Beating common sense into interactive applications. *AI Magazine*, 25(4):63–76, 2004.
- [23] Hugo Liu and Push Singh. Commonsense reasoning in and over natural language. In *KES*, pages 293–306, 2004.
- [24] D.L. McGuinness, F. Van Harmelen, et al. Owl web ontology language overview. Technical Report 2004–03, W3C recommendation, 2004.
- [25] Marvin Minsky. Deep issues: commonsense-based interfaces. *Commun. ACM*, 43(8):66–73, 2000.
- [26] Jorge Pérez, Marcelo Arenas, and Claudio Gutierrez. nsparql: A navigational language for rdf. *J. Web Sem.*, 8(4):255–270, 2010.
- [27] CLASS project. <http://class.inrialpes.fr/>. Technical report, 2006.
- [28] Eric Prud’hommeaux and Andy Seaborne. Sparql query language for rdf. Technical report, W3C Recommendation, January 2008.
- [29] A. Pugliese, O. Udrea, and VS Subrahmanian. Scaling RDF with time. In *Proceeding of the 17th international conference on World Wide Web*, pages 605–614. ACM, 2008.
- [30] N. Ramakrishnan, D. Kumar, B. Mishra, M. Potts, and R.F. Helm. Turning CARTwheels: an alternating algorithm for mining redescriptions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 266–275. ACM, 2004.
- [31] R. Ramakrishnan and J.D. Ullman. A survey of deductive database systems. *The journal of logic programming*, 23(2):125–149, 1995.
- [32] Entity resolution. <http://www.cs.umd.edu/projects/linqs/projects/er/index.html>. Technical report, 2008.
- [33] Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public. In *CoopIS/DOA/ODBASE*, pages 1223–1237, 2002.
- [34] Umberto Straccia, Nuno Lopes, Gergely Lukacsy, and Axel Polleres. A general framework for representing and reasoning with annotated semantic web data. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.
- [35] D. Turnbull, R. Liu, L. Barrington, and G. Lanckriet. A game-based approach for collecting semantic annotation of music. In *Proceedings of the International Symposium on Music Information Retrieval*, 2007.
- [36] K. Verboven, M. Carlier, and J. Dumolyn. A short manual to the art of prosopography. In KEATS-ROHAN Katharine S. B., editor, *Prosopography Approaches and Applications. A Handbook*, pages 35–70. Prosopographica et Genealogica, Oxford, Unit for Prosopographical Research, 2007.
- [37] J. Whaley, D. Avots, M. Carbin, and M. Lam. Using datalog with binary decision diagrams for program analysis. *Programming Languages and Systems*, pages 97–118, 2005.
- [38] S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *Principles of Data Mining and Knowledge Discovery*, pages 78–87. Springer, 1997.

- [39] M.J. Zaki and N. Ramakrishnan. Reasoning about sets using redescription mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 364–373. ACM, 2005.

5 publications of the applicants

- Geert Jan Bex, Frank Neven, Thomas Schwentick, **Karl Tuyls**: Inference of Concise DTDs from XML Data. *VLDB*: 115-126 (2006)
- Steven de Jong, **Karl Tuyls**: Human-inspired computational fairness. *Autonomous Agents and Multi-Agent Systems* 22(1): 103-126 (2011)
- **Toon Calders** and Bart Goethals: Non-derivable itemset mining. In: *Data Min. Knowl. Discov.* Vol. 14(1): pp. 171-206 (2007)
- **Toon Calders**, Calin Garboni, and Bart Goethals: Approximating Frequentness Probability of Itemsets in Uncertain Data. In: *Proceedings IEEE ICDM International Conference on Data Mining* (2010)
- Saso Dzeroski, Luc De Raedt, **Kurt Driessens**: Relational reinforcement learning. *Machine Learning*, Vol. 43: 7–52 (2001)