系図からのデータ自動取得の試み

永井 謙也　村川 猛彦（和歌山大学）
大澤 留次郎（凸版印刷）
宇都宮　啓吾（大阪大谷大学）

系図には人物の詳細や人間関係などが記載されており，人文学の研究において有益な資料である．従来の系図を対象としたシステムにおけるデータの入力は支援があるものの手動で行っていた．本研究では，位置情報付きテキストとして表現されている系図データに対し，人物名と付随情報の判別，人物同士ならびに人物と付随情報の関連付けの自動化を試みた．判別にあたっては，人物名と付随情報で高さと幅が異なる点や，人物名における特徴的な語に着目した．また系図画像上で縦横の線分が明瞭に描かれており，画像内の文字列領域は位置情報付きテキストの座標情報をもとに削除できることを用いて，線分検出を行った．判別の成功率は 90%を超え，4 枚の系図画像に出現する 75 人の兄弟関係などの自動取得ができた．

An attempt of automatic data extraction from genealogies

Kenya Nagai　Takehiko Murakawa (Wakayama University)
Tomejiro Osawa (Toppan Printing)
Keigo Utsunomiya (Osaka Ohtani University)

Genealogies are useful materials in human literature research since they contain the details of persons and relationships among persons. When constructing genealogical database systems, the data have been registered manually although there was support. In this research, we attempted to automate the discrimination between personal names and supplementary notes, the acquisition of the human relationships, and the association of persons with notes, using text files with location information extracted from genealogical image files. Moreover we implemented the detection of horizontal and vertical line segments displayed on genealogical images. As a result of applying to a couple of genealogies, the success rate of the discrimination exceeded 90%, while we found a sibling group of 75 persons that appear on four image files.

１．はじめに

系図には人物の詳細や人間関係などが記載されており，人文学の研究において有益な資料である．一部の系図は画像化されており，インターネット上に公開されている．これらは画像の閲覧しかできないため必要な情報を得るには，それを探す時間と手間がかかってしまう．これらの問題を解決するために系図を対象としたデータベースシステムの開発が行われてきた．

筆者らはこれまで，平安・鎌倉時代を中心とする僧侶の師弟関係を表した系図の情報をデータベースに格納し，Java のツリーマップを用いて関係を図示することで，系図の検索や閲覧ができるインタフェースを構築してきた[1]．また山本ら[2]は，加茂祢宜神主系図を対象としたデータベースの構築および運用について報告している．そこでは位官職や父子関係などの人物の情報と叙位や任官などの年月日の情報を分けてデータベース管理し，人名，官職名の検索だけでなく年表の検索もできるようになっており，全ての記載人物に対して系図画像のリンクが貼られている．相田[3]の系図データベースシステムは日本古典系図を対象としている．データテーブルの作成を 2 段階に分けており，1 段階目では人物に ID を振り，簡略な情報付加を手作業で行う．2 段階目では 1 段階目に作成したテーブルの情報をもとに，人物マスターテーブルと細目マスターテーブルを作成している．人物マスターテーブルには名前や親子関係や兄弟関係を，細目マスターテーブルには年月日，官職歴の情報を管理している．親子関係や兄弟関係は 1 段階目に作成したテーブルの情報から自動で識別する処理を行っている．

このように系図を対象とした検索や閲覧ができるシステムの研究や開発は幅広く行われているが，このようなシステムにおけるデータの入力は，ある程度の支援はあるものの手動で行っているのが現状であり，系図データベースの構築には時間と手間がかかっていた．

そこで本研究では，位置情報付きテキストとし

て表現された系図データに対し，人物名と付随情報の識別，人物同士ならびに人物と付随情報の関連付けの自動化を試み，もとの系図画像を利用して，系図の検索や閲覧が可能なデータベースシステムの開発を行ってきた．このシステムの活用によって，さまざまな系図データをより効率良くデータベース化し閲覧，調査できるシステムの開発を目指している．

２．対象とする系図およびデータ

２．１　系図

系図とは一族の代々の系統を書き表した図表のことである．血脈関係のみならず，学芸の師匠から弟子への師資相承の関係（師弟関係）を表した図表も本研究の対象である．本稿では師弟関係でも親子関係，兄弟関係と表記する．また対象とする系図には婚姻関係の記載はされていない．系図は撮影画像，出版物，文書ファイルなどの形態で表現されている．

系図に書かれる文字列は人物名と付随情報に大別される．図 1 は将軍家譜[4]の系図の撮影画像であり秀吉，秀長，女子が人物名にあたる．付随情報とは系図や人物の詳細な情報のことであり図 1 では不詳其父などが該当する．

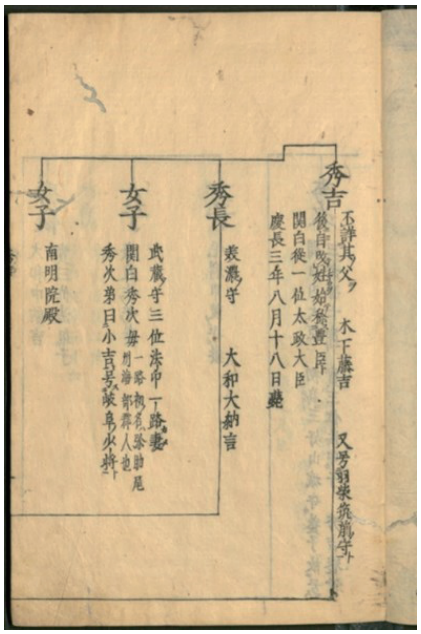


図 1　将軍家譜の系図画像例

出版物では，例えば図 2 のような系図がある（真言宗付法血脈[5]）．この系図は，原本をもとに翻刻者が読みやすくし，また必要に応じて情報を付記している．権律師真然，聖寶などが人物名，高野籠山，授法弟子三十五人などが付随情報にあたる．

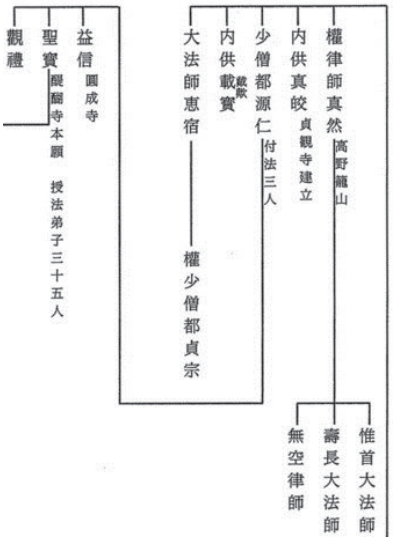


図 2　真言宗付法血脈の系図記載例

これらの系図，および次節で述べる系図を参照して得られた共通点を整理する．まず，人物名の近くにその人物の付随情報が記載されており，人物の親子関係や兄弟関係は，縦と横の線分の組み合わせでつながっている．次に，特筆すべきことは，人物から伸びる線分は必ず縦の線分となっている点である．これらの特性を考慮することで，自動処理によって漏れなく短時間で情報の抽出ができると考えられる．

２．２．位置情報付きテキスト

本研究では系図画像（系図が出版物，または PDF 形式などの文書ファイルの場合には，画像に変換しておく）に加え位置情報付きテキストを使用する．位置情報付きテキストは XML 文書であり，人物名もしくは付随情報の文字列情報に，画像上の位置を表す左上 XY 座標と，文字列の大きさを表す高さ・幅の 4 つの情報が記述される．例えば図 1 の秀吉については，「<String CONTENT=" 秀 吉 " HEIGHT="260" HPOS="1641" VPOS="809" WIDTH= "157">」となり，左上座標は(1641,809)，幅と高さはそれぞれ 157 と 260 と読み取ることができる．位置情報付きテキストの作成には，凸版印刷の OCR ソフトウェアを使用した．ただし，このテキストでは人物名と付随情報の区別がなされておらず，線分の情報は記載されていない．

位置情報付きテキストの作成を試みた系図の一部を表 1 に示す．位置情報付きテキストのファイル数は画像枚数と一致するため省略した．また画像サイズは同一系図においても若干の違いが見られる．

表 1 位置情報付きテキストを作成した系図例

系図名および出典	画像枚数	画像サイズ(px)
将軍家譜[4]	11	2024x3239
真言宗付法血脈[5]	13	2358x1715
東寺観智院金剛蔵本『真言付法血脈　仁和寺』[6]	38	2146x3000
東寺真言宗血脈[7]	18	1827x2495
天台血脈[8]	51	2154x3034
東寺観智院金剛蔵本『密教師資付法次第　千心』[9]	39	2154x3034
醍醐寺本『伝法潅頂師資相承血脈』[10]	75	2142x3000
仁和寺蔵『真言伝法潅頂師資相承血脈』[11]	83	2480x3479
真言相承諸流血脈図[12]	25	1742x2438

3．データ取得

系図画像および位置情報付きテキストを所与とし，その情報を効率良く活用するため，図 3 に示す処理の流れにより，情報の抽出を試みた．最終的にデータベースに格納するのは，「付随情報」と「人物」とし，線分座標は一時的なファイルに保存するが，データベースには格納していない．以降では，「人物名と付随情報の判別」「人物名と付随情報の対応づけ」「線分検出」「人物同士の対応付け」の順に，詳しく説明していく．

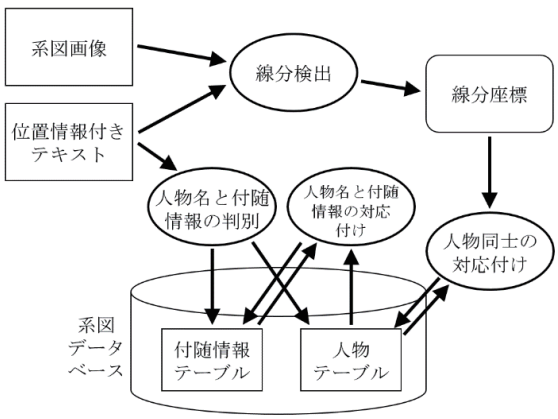


図 3　データと処理の流れ

3．1．人物名と付随情報の判別

将軍家譜に記載された系図画像 11 枚分の位置情報付きテキストのデータ（文字列の総数 198，人物名 69，付随情報 128）をもとにして，文字列ごとに高さ，幅，X 座標，Y 座標の値を調べ，人物名と付随情報の判別に有効な条件を調査した．

判別方法を決める前に，系図画像を参照しながら目視で人物名と付随情報の判別を行い，高さと幅について散布図を作成した（図 4）．人物名が散布図の右下に多く分布していることから，人物名は付随情報に比べ文字の幅が大きく文字の高さが小さいものが多いことが分かった．また文字列の高さ，幅の平均値は人物名と付随情報を分ける閾値に近くなっており，そこから人物名と付随情報を判別できるのではないかと考えた．

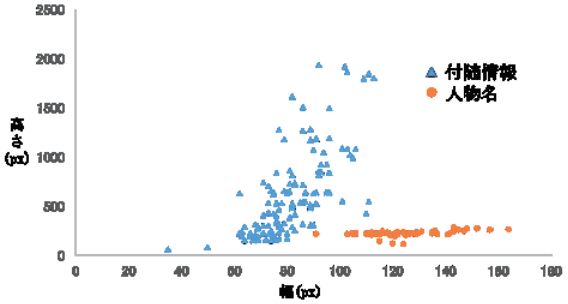


図 4　将軍家譜の位置情報付きテキストに出現する文字列の高さと幅の散布図

3．2．人物名と付随情報の対応付け

各文字列領域の左上を基準として，X 座標，Y 座標の値を用いて付随情報から人物名までの距離をそれぞれ計算し，その距離が近い人物名と付随情報を対応付ける処理を行う．

1 人の人物名に対し複数の付随情報が記載されているために，付随情報とそれに対応する人物名との距離が遠くなってしまい，他の人物名の付随情報として誤判定される可能性がある．多くの系図の付随情報は横一列，または縦一列に並んでいることを利用し，並んでいる付随情報をグループにすることで人物名と対応付けを行う．

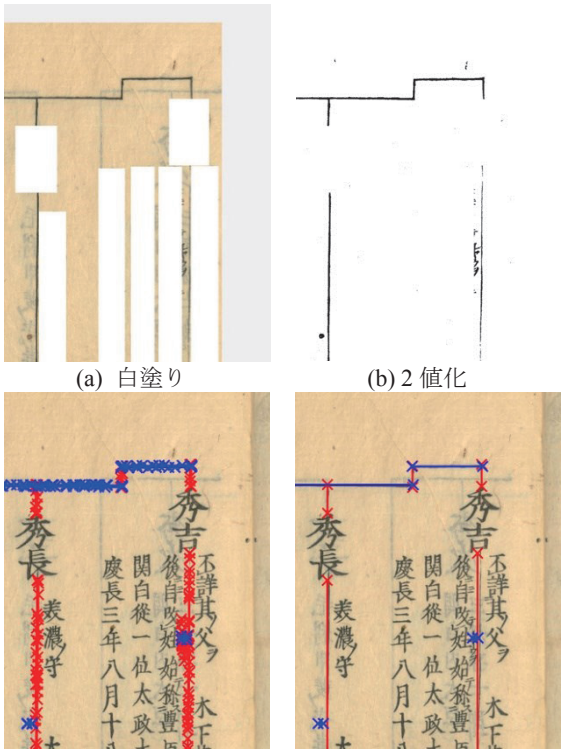
3．3．線分検出

系図画像に出現する，人物同士を結ぶ線分の検出を試みた．系図画像において，縦横の線分が明瞭に描かれていること，および画像内の文字列の領域は，維持情報付きテキストの座標情報をもとに削除できることなどを考慮し，独自に実装した．

手順の概略を述べる．まず，検出対象外の領域および文字列領域に該当する画像の箇所を白で埋めてから，グレイスケール化ののち 2 値化を行う．次にピクセル単位で，行方向・列方向に黒の画素を走査して線分の候補を取得する．そして隣接する候補を統合して横方向・縦方向の線分を求める．

この処理において，(1)検出対象の領域を表す矩形座標（左上と右下の XY 座標），(2) 2 値化の閾値，(3)線分の候補の最小連続長，および(4)同方向の線分同士を統合するための距離の上限は，処

理対象となる画像により値が異なる．これらの最適値の算出も可能と考えられるが，本研究では様々に値を変えて実行し，結果を画像上に描画して，目視により，良いものを選んだ．図 1 の画像を対象とした描画例を図 5 に示す．図 5 (c)-(d)では線分の両端に×印を描いて強調している．



(a) 白塗り　(b) 2 値化

(c) 線分の候補　(d) 線分を統合

図 5　線分認識例

線分検出の処理は，プログラミング言語 Ruby を用いて開発した．2 値化までの処理には画像処理ライブラリの Netpbm を用い，得られた ppm 形式のファイルを開いて画素を読み出した．

3．4．人物同士の対応付け

系図において人物同士の親子関係，兄弟関係は線分のつながりによって記載されている．例えば人物名 P について，その下の縦線から到達できる人物名 Q は P の子である．また人物名 P について，その上の縦線から到達できる人物名 Q は P の親または兄弟となる．いずれも，一つの P に対して Q は複数存在し得る．

人物名 P，Q が親子であるかを判別するためには，人物名 P の下にのびる縦線が人物名 Q の上にのびる縦線とつながっているかを調べればよい．人物名の座標と 3.2 節で求めた線分の座標から線分の先と人物名の対応付けを行い，人物名の下にのびる線分が他の人物名の上にのびる線分と同一またはつながっている場合，その人物同士は親子関係であると分かる．このとき，線分同士が交差しているかどうかの判定を行うことによって，複数の線分によって記載されている親子関係も獲得できる．

人物名の上の縦線から到達できる人物名は親または兄弟のどちらかである．兄弟関係を判別するためには，その親，兄弟から親を取り除くことで求めることができる．図 1 の系図上では秀吉，女子，女子の親が記載されていないが，兄弟の関係であり，上記のルールで兄弟関係を獲得することが可能となる（親子関係をもとに，同一の親を持つ者を兄弟とする，というルールでは得られない）．

系図によっては，人物同士の関係が複数の系図画像にまたがって記載されていることがある．その場合は連続する系図画像の横線分を照合して（複数あれば上から順に）連結することとした．これにより複数の系図に人物の関係が分かれて記載されていても問題なく人物同士の関係を対応付けることができる．

ここで「つながり」の基本となる，線分交差の判定方法について述べる．縦線・横線の交差の形状として，L 字型，T 字型，十字型が考えられるが，本研究では「線につながりがあるか」を判断できればよいので，形状を問わず判定できることとした．また，前節の処理によって，先端部分のかすれのため線分が短めに認識された場合でも，つながるものとしたい．

座標を用いて，判定方法を説明する．判定対象の縦線の上下端の座標を(x_1, y_{11})および(x_1, y_{12})，横線の左右端の座標を(x_{21}, y_2)および(x_{22}, y_2)とする．ただし$y_{11} < y_{12}$，$x_{21} < x_{22}$である．ここで 2 つの線分の各両端を，それぞれ定数τだけ伸ばす．得られる縦線の上下端の座標を(x_1, y'_{11})および(x_1, y'_{12})，横線の左右端の座標を(x'_{21}, y_2)および(x'_{22}, y_2)と表したとき，$y'_{11} = y_{11} - \tau$，$y'_{12} = y_{12} + \tau$，$x'_{21} = x_{21} - \tau$，$x'_{22} = x_{22} + \tau$となる．そして，$x'_{21} \le x_1 \le x'_{22}$および$y'_{11} \le y_2 \le y'_{12}$の両方が成り立つとき，この縦線と横線は交差すると判断する．L 字型の交差例を図 6 に示す．

図 6　L 字型交差の座標例

十字型の交差を検出したとき，つながっていると見なすべきか，それとも描画の都合で交差しているがつながっていないと見なすべきかについては，個別に判断しなければならない．将軍家譜には十字型の交差は見られなかったが，真言宗付法血脈では，文献[5]の p.441 に，1 種類ずつ出現し，他のページにも現れる．この系図においては，十字型で交差する横線の両端が，それぞれ他の縦線の上端と L 字型で交差している場合のみ，つながっていると判断すればよい．しかし他の系図でもこのルールで判断できるかどうかは不明である．

4．判別結果

3 節で提案した人物名と付随情報の判別，人物名と付随情報の対応付け，人物同士の対応付けを，いくつかの系図に対して行った．

4．1．人物名と付随情報の判別結果

将軍家譜の系図に記載された文字列 198 個の平均値である高さ 464 以下かつ幅 96 以上の文字列を人物名と判定しその結果を目視で確認した（図 7）．その結果，69 個の人物名のうち 68 個の人物名を人物名であると判定し識別率は 98.5%となった．また付随情報の 1 つが人物名と誤検出される結果になった（表 2）．なお，付随情報は複数行になることもあるが，位置情報付きテキストの記載に基づき，分割・統合せずに数えている．

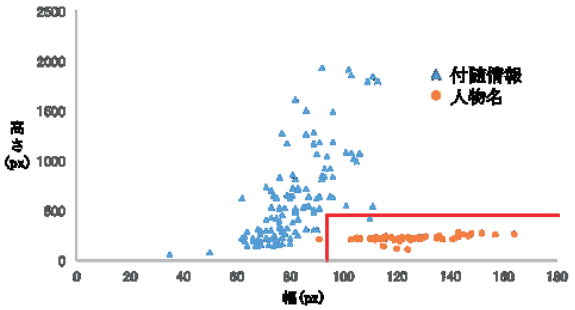


図 7　将軍家譜の位置情報付きテキストに出現する文字列の高さと幅の散布図

これに対して系図の人物名と付随情報はそれぞれ縦または横に一直線上に並んでいる性質を利用し，人物名と付随情報をグループに分け，もう一度判別を行ったところ，人物名 60 個すべてを人物名と識別することができ，付随情報が人物名と誤検出されることもなかった．

表 2　将軍家譜の人物名と付随情報の判別結果

	総数	識別成功	識別失敗
人物名	69 個	68 個	1 個
付随情報	128 個	127 個	1 個

多くの系図は将軍家譜のように人物名のほうが付随情報より文字の大きさが大きく，また付随情報は人物名に比べ文字数が多くなっているため，他の系図に対しても本研究の判別手法は使用できると考えている．しかし一部の系図では人物名と付随情報に明確な大きさの区分がなく，人物名のほうが付随情報よりも文字数が多くなっているものがある．この特徴は僧侶の師弟関係の系図に見られ，真言宗付法血脈もそのひとつである．

真言宗付法血脈に記載されている文字列 870 個において散布図を作成したところ図 8 のようになった．将軍家譜と違い文字の大きさでは識別できないため，人物名と付随情報の判別には別の手法をとる必要がある．

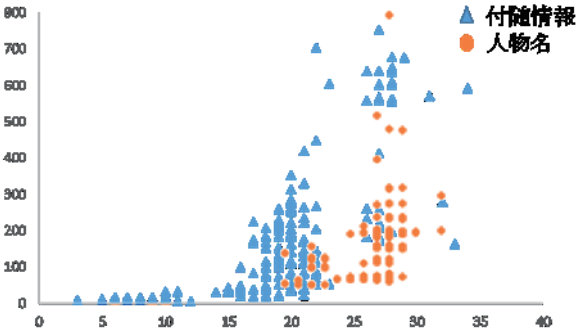


図 8　真言宗付法血脈の位置情報付きテキストに出現する文字列の高さと幅の散布図

真言宗付法血脈の人物名に着目すると，法師，律師，阿闍梨といった僧の敬称や僧侶を表す名が多く使われていることが分かった．これらの数を数えると人物名 449 個に対し僧正，僧都，法師，阿闍梨，律師と名のつく人物名は 371 個記載されており，人物全体の 8 割以上を占めた．

真言宗付法血脈も将軍家譜と同じように人物名と付随情報はそれぞれ縦または横に一直線上に並んで記載されている．このことから僧正，僧都，法師，阿闍梨，律師と名のつく文字を人物名とし，その文字列と縦並び，横並びになっている人物名をグループにすることによって人物名を判別できるのではないかと考えた．この処理を行い，人物名を識別したところ人物名 449 個のうち 434 個の人物名を適切に識別することが出来た．しかし付随情報であるのに人物名と識別された文字が 22 個に上った（表 3）．

表 3　真言宗譜法血脈の人物名と付随情報の判別結果

	総数	識別成功	識別失敗
人物名	449 個	434 個	15 個
付随情報	421 個	399 個	22 個

東寺観智院金剛蔵本『真言付法血脉　仁和寺』[6]については，将軍家譜と同様に，高さがこの系図の文字列全体の平均値以下かつ幅が平均値以上を抽出し，その中で丸囲み数字やカッコ類などを含むものを除外した文字列を，人物名としたとき，人物名と付随情報を合わせた識別率は，3533／3545＝99.7%となった．誤判別のうち 1 個は，人物名を付随情報としたものであり，残り 11 個は，字数の少ない付随情報を人物名としたものであった．

4．2．人物名と付随情報の対応付けの結果

将軍家譜の系図 11 枚において人物名と付随情報の判別を行った後，それぞれを対応付ける処理を行った．人物名と付随情報の距離が近いものを対応づける処理を行ったところ付随情報 128 個のうち 127 個を適切な人物名と対応付けることが出来た．うまく対応付けが出来なかった 1 つは人物名に多くの付随情報が記載されており，対応する人物名との距離が遠くなってしまっていたためであり，横並びになっている付随情報をグループにする処理を行うことで適切に対応付けが行えた．

4．3．線分検出および人物同士の対応付けの結果

将軍家譜の 11 枚の画像を対象に，3.2 節で述べた処理を行った結果，77 本の横線および 145 本の縦線を検出した．図 5 (d)のように描画した画像を見ながら，実態に合うよう線分の追加や削除，統合を行った結果，48 本の横線および 101 本の縦線となった．真言宗付法血脈の 13 枚の画像に対しては，自動処理では 140 本の横線と 557 本の縦線が得られ，人手により 136 本の横線と 567 本の縦線となった．

横線・縦線の交差（L 字型，T 字型，十字型のすべて）は，将軍家譜で 102 箇所，真言宗付法血脈で 589 箇所となった．

2 種類の系図のそれぞれで，人物名と縦線との対応付け，およびページをまたいだ横線同士の対応付けを，実装した処理に基づき行ったのち，人物同士の親子関係・兄弟関係の導出を試みた．将軍家譜では 59 組，真言宗付法血脈では 496 組の親子関係を見つけることができた．兄弟関係のグループは，将軍家譜では 11，真言宗付法血脈では 44 あり，真言宗付法血脈で 4 枚の連続する画像にまたがる，75 名からなる兄弟関係も取得できている．

図 2 の画像について，認識した線分，交点，人物名および付随情報の箇所に色づけしたものを，図 9 に示す．

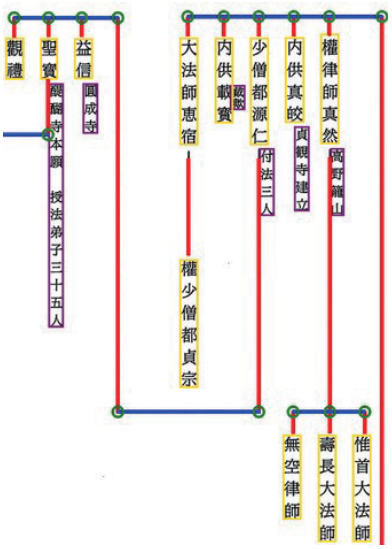


図 9　真言宗付法血脈の認識例

5．データベース構築およびデータ活用

系図データおよび上記の処理で取得できる情報を効率良く活用するため，データベースに情報を格納する．その設計・構築にあたり，「系図文献」「系図画像」「人物」「付随情報」について 1 対多対応が得られることなどから，関係データベースを採用し，それぞれをテーブルとした．

位置情報付きテキストから得られる文字列や座標は，人物または付随情報の属性とした．親子関係は，人物テーブルの自己参照キーにより表現できる．この関係をまとめた ER 図（実体関連ダイアグラム）の主要部を図 10 に示す．3.4 節で述べた，親のない兄弟関係については，それらの親となるダミーのレコードを人物テーブルに登録し，その人物 ID には負数で兄弟関係ごとに異なるものを割り当てている．

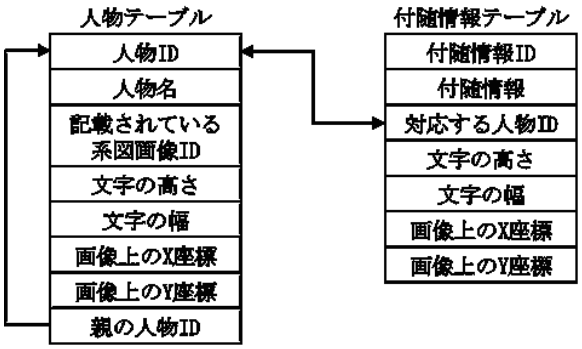


図 10　系図データベースの ER 図

このデータベースを活用した系図の検索・閲覧のできる系図システムのインタフェースを開発している．このシステムではこれまで開発を行ってきた[1]のような独自の系図の見せ方をするインタフェースを構築するのではなく，もとの系図画像を表示しその上に情報を配置している．マウスカーソルで系図上の文字をクリックすると画面右側にその文字を表示することができる．これによってもとの画像を拡大しても文字がつぶれてしまって見えない小さな字などを見ることができる．図 11 では真言宗付法血脈の系図画像に位置する人物名の僧正真濟をマウスでクリックした様子である．人物名をクリックした場合は，その人物の付随情報が下に表示される仕組みになっており，僧正真濟をクリックすると高尾が表示される．

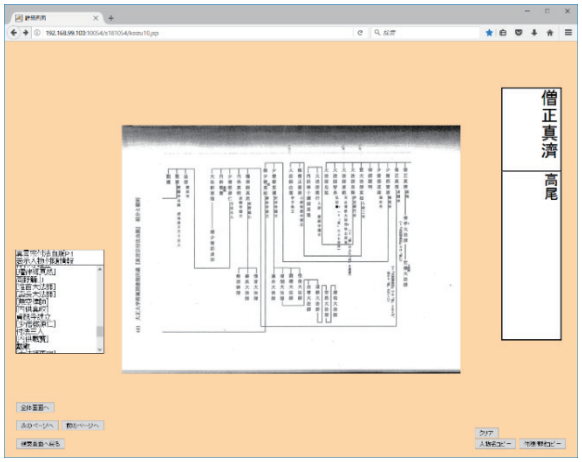


図 11　系図閲覧インタフェース例
（系図詳細画面）

系図によっては複数の画像によって人物関係が記載されているため，全体像を把握するためには一目で系図全体を見られる機能があるとよい．そこで図 12 のように系図全体を横にスクロールすることで閲覧できる機能を作成した．系図はマウスでドラッグすることで自由に移動させることができ，手に持って系図を見るのと同じ感覚で系図の閲覧ができる．これによってページをまたいだ人物関係の参照がしやすくなる．

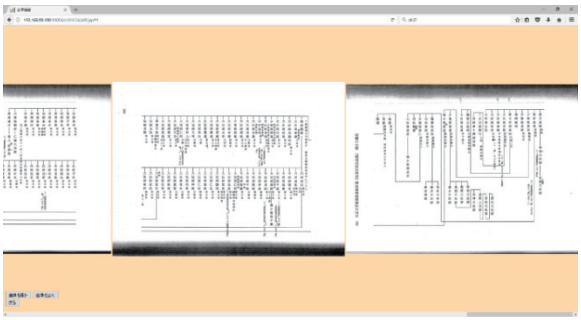


図 12　系図閲覧インタフェース例
（系図一覧画面）

６．おわりに

系図を対象としたシステムにおけるデータベースの構築において従来ではデータを手入力するしかなく構築に時間と手間がかかっていた問題を解決するために，系図の情報を自動で取得し，適切な情報の対応付けを行ってデータベースに格納するシステムの構築を行った．人物名と付随情報の判別には少し誤検出が見られたが，人物名と付随情報の対応付けと人物同士の関係については高精度な自動判別を行えるシステムになっている．誤検出に関しては系図によって少なからず生じるものであるため，人手による手軽な確認および修正の作業をこのシステムの中で行えればよいと考えている．そのためのシステム改良が今後の課題になる．

謝辞　系図閲覧インタフェースの開発にあたっては村端宏介氏の協力を得ました．ここに記して感謝申し上げます．本研究は JSPS 科研費 JP26284065，JP17H02342 の助成を受けたものです．

参考文献

[1] 田中猛彦，冨金原賢次，宇都宮啓吾，中川優：平安・鎌倉を対象とした僧侶データベースシステム，情報知識学会誌，Vol.13，No.2，pp.18-31 (2003).
[2] 山本宗尚，月本一武：『賀茂祢宜神主系図』データベースの構築と活用の可能性，じんもんこん 2015 論文集，pp.203-210 (2015).
[3] 相田満，日本古典データベースの構築，情報処理学会研究報告 人文科学とコンピュータ，2001-CH-051，pp.39-46 (2001).
[4] 人文学オープンデータ共同利用センター準備室日本古典データセット　将軍家譜.

http://codh.rois.ac.jp/pmjt/book/200021823/ (2017-11-15 参照).
[5] 苫米地誠一, 大正大学附属図書館所蔵『真言宗付法血脈』紹介と翻刻, 川勝守・賢亮博士古稀記念東方学論集, 汲古書院, pp.431-452 (2013).
[6] 武内孝善, 東寺観智院金剛蔵本『真言付法血脈仁和寺』, 高野山大学密教文化研究所紀要通号 6, pp.39-131 (1993).
[7] 坂本正仁, 東密血脈譜叢刊（一）東寺真言宗血脈, 豊山学報通号 31, pp. 27-52 (1986).
[8] 武内孝善, 東寺観智院蔵『天台血脈』の研究（一）, 高野山大学論叢通号 39, pp. 13-87 (2004).
[9] 武内孝善, 東寺観智院金剛蔵本『密教師資付法次第千心』, 高野山大学論叢通号 28, pp.187-222 (1993).
[10] 築島裕, 醍醐寺蔵本「伝法灌頂師資相承血脈」解題, 醍醐寺文化財研究所研究紀要通号 1, pp.27-135 (1978).
[11] 阿部泰郎編, 横山和弘, 佐藤愛弓執筆, 真言伝法灌頂師資相承血脈, 名古屋大学比較人文学研究年報第 4 集 (2003).
[12] 坂本正仁, 東密血脈譜叢刊二　真言相承諸流血脈図, 豊山学報通号 40, pp.41-79 (1997).