International Journal on Document Analysis and Recognition (IJDAR) (2023) 26:255–272
https://doi.org/10.1007/s10032-023-00427-w

SPECIAL ISSUE PAPER

Large-scale genealogical information extraction from handwritten Quebec parish records

Solène Tarride[1] · Martin Maarand[1] · Mélodie Boillet[1,2] · James McGrath[3] · Eugénie Capel[3] · Hélène Vézina[3] · Christopher Kermorvant[1,2]

Received: 10 November 2022 / Revised: 10 November 2022 / Accepted: 10 January 2023 / Published online: 30 January 2023

Abstract
This paper presents a complete workflow designed for extracting information from Quebec handwritten parish registers. The acts in these documents contain individual and family information highly valuable for genetic, demographic and social studies of the Quebec population. From an image of parish records, our workflow is able to identify the acts and extract personal information. The workflow is divided into successive steps: page classification, text line detection, handwritten text recognition, named entity recognition and act detection and classification. For all these steps, different machine learning models are compared. Once the information is extracted, validation rules designed by experts are then applied to standardize the extracted information and ensure its consistency with the type of act (birth, marriage and death). This validation step is able to reject records that are considered invalid or merged. The full workflow has been used to process over two million pages of Quebec parish registers from the 19–20th centuries. On a sample comprising 65% of registers, 3.2 million acts were recognized. Verification of the birth and death acts from this sample shows that 74% of them are considered complete and valid. These records will be integrated into the BALSAC database and linked together to recreate family and genealogical relations at large scale.

Keywords Information extraction · Document layout analysis · Handwritten text recognition · Historical documents · Quebec parish records

1 The BALSAC project

1.1 From BALSAC to i-BALSAC

For the last 50 years, the BALSAC project[1] has been building and consolidating a major database on the Quebec population. The core of the database is made of demographic events extracted from transcribed parish and civil registers. Birth, marriage and death records are linked together to reconstruct the Quebec population from the beginning of French settlement in the seventeenth century to the contemporary period. From the 1980s, mostly marriage records were entered in the database, in order to concentrate on genealogical reconstructions used in genetic research.

About ten years ago, the decision was made to add birth and death records and link them to the marriages already in BALSAC for a comprehensive coverage of families. During this time, it also became increasingly evident that the development of the database, which entails work on millions of records, could no longer rely exclusively on manual or semi-automatic operations. Fortunately, progress in machine learning opens up promising avenues for historical databases as handwritten text recognition (HTR) algorithms have improved significantly in the past few years. These considerations have led the BALSAC team to make the decision to rely on this technology for the transcrip-

[1] https://balsac.uqac.ca/

✉ Solène Tarride
starride@teklia.com

Mélodie Boillet
boillet@teklia.com

Christopher Kermorvant
kermorvant@teklia.com

[1] TEKLIA, Paris, France

[2] LITIS, Normandie University, Rouen, France

[3] BALSAC project, Université du Québec à Chicoutimi, Saguenay, Canada