This page was exported from - Digital meets Culture
Export date: Wed Nov 18 17:44:01 2026 / +0000 GMT

From Archive to Dataset: AI-Assisted Mass Digitisation of Historical Photographic Collections, Using the Sven Simon Archive as a Case Study



Saint Joan of las Abadesas' archive. Provided by Ajuntament de Girona.

Paper for the 19th Image and Research Conference in Girona, 19-20 November 2026.

Abstract]
Historical photographic archives contain considerably more information than the comparatively small number of images traditionally selected for digitisation and publication. United Archives is therefore developing a scalable process that combines high-throughput scanning with AI-assisted metadata creation in order to digitise complete analogue photographic archives as comprehensively as possible. The objective is not merely to preserve individual images, but to transform entire collection contexts into structured, searchable datasets. The first large-scale implementation is the Sven Simon Archive, comprising approximately 1.6 million photographs relating to German and international political, cultural, economic and sporting history since the early 1960s.

The paper describes the collaboration with a specialised scanning company in Cologne, the linking of physical objects to digital files, the capture of inherited captions and the use of multimodal AI to produce structured metadata drafts. Experience to date shows that AI substantially accelerates descriptions, keywords, translations and the structuring of heterogeneous source information. It does not, however, replace archival contextual knowledge or the verification of people, places, dates, events and rights.

Comprehensive digitisation is particularly valuable because it reveals series, developments over time and previously overlooked relationships, while enabling new forms of data use in journalism, research, education, historical analysis and future AI applications. Prerequisites include traceable metadata provenance, risk-based quality assurance and a clear separation between source evidence, machine inference and human-verified statements.

Keywords: photographic archives; comprehensive digitisation; artificial intelligence; metadata; Sven Simon; mass scanning; data value; historical analysis.

Read the full paper<<<<