Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 314 (2015) 83–100

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

Software Architecture for Document Anonymization

Horacio Vico[1]

División Informática
Poder Judicial
Montevideo, Uruguay

Daniel Calegari[2]

Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay

Abstract

Organizations often have a dilemma in relation to their documents: ensure confidentiality of the data or publish the information contained in them, for transparency, scientific interest, or other reasons. In this context is that document anonymization arises, i.e. the replacement of sensitive data in such a way that preserves the confidentiality of the documents without altering their value or usefulness. There are proposals for (semi)automatic anonymization, but they are often domain-specific or they partially address the problem. In this paper we present a software architecture for supporting document anonymization, which is based on the representation of the problem as a domain and platform independent configurable business process. In addition, we analyze the technological alternatives for implementing the architecture and we present a functional prototype applied to the domain of legal documents.

Keywords: document anonymization, software architecture, business process

1 Introduction

Document management is the set of activities for the creation, reception, organization, storage, preservation, access and dissemination of documents within an organization [20]. Through the use of technology it is possible to improve management and maximize the value of the huge amount of information within those documents. In this context, two apparently conflicting interests arise: organizations must ensure confidentiality of the personal information they handle, but restricting access to such information is not a valid alternative. This happens whenever the

[1] Email: hvico@poderjudicial.gub.uy
[2] Email: dcalegar@fing.edu.uy

http://dx.doi.org/10.1016/j.entcs.2015.05.006

organization needs to make available much of their information, either because of transparency or because the stored documents have a scientific, biomedical or legal value which has no relation with the personal information contained within them.

A solution to this problem is document anonymization, i.e. the total or partial replacement of personal data by eliminating any reference to their identity, without altering the value or usefulness of the original document [6]. Such anonymization can be manually performed by a user, being tedious, repetitive, error-prone and time consuming, not providing added value to the organization. This problem has driven research and development of techniques and methodologies for (semi)automatic anonymization.

The problem is not trivial given that many documents have no structured format to easily identify the sensitive information within them, thus it is necessary to combine different computational disciplines such as natural language processing, text mining, and machine learning to solve the problem. Particularly, from the point of view of the software architecture, the integration of different technological elements that can be used in an anonymization process represents a major challenge.

There are previous works proposing architectures to tackle with this problem [16,12,9], but they have some limitations: (a) they are domain-specific solutions making them hard to adapt to other contexts; (b) they partially resolve the problem without clearly defining the overall anonymization process; (c) they are not flexible enough to add new tools into the process; (d) the implementation of these solutions is not public and therefore it is not possible to experiment with them.

In this paper we present a software architecture that supports document anonymization overcoming the limitations of existing architectures. In particular, this architecture represents the problem as a domain and platform independent configurable business process [23]. This allows its implementation using a Business Process Management System (BPMS, [23]). Furthermore, we analyze different technological alternatives to implement a functional version of the reference architecture with freely available tools, thereby reducing licensing fees. Finally, we made a more qualitative assessment of the proposal by implementing a functional prototype applied to the domain of legal documents, in the context of the Judiciary of Uruguay.

The rest of this paper is structured as follows. In Section 2 we introduce the main aspects concerning document anonymization. In Section 3 we present the main architectural initiatives addressing this problem and we describe their common and special features. In Section 4 we describe our proposal and in Section 5 we present the development of a prototype of the reference architecture and its application on a case study. Finally, in Section 6 we present the main conclusions and some ideas for future work.

2 Anonymization of Documents

Very often, documents stored in an organization contain personal or sensitive information of citizens or legal persons, whose privacy must be guaranteed by the