

Proposed 10/22/08

# EGov: Canoncial PDF Text Extraction Algorithm

Signaturwert	sQnb04p1ymGHRBpowty7NrmasysWwBR8ATwEtS/xDyM3SQ3xNLB8Q4K+2t0H2vU	
Unterzeichner	C=AT, OU=Vsig, O=Hauptverband Österr. Sozialvers., CN=Thomas Rössler	
Datum/Zeit-UTC	2007-06-25T11:39:51Z	
Aussteller-Zertifikat	C=AT, O=Hauptverband Österr. Sozialvers., CN=Vsig CA 2	
Serien-Nr.	17176797848875370451084887172968614235987	
Methode	urn:pdfsigfilter:bka.gv.at:text:vi.0.0	
Parameter	ets1-bka-1.0@1192771532-199060984480-7742-20016-7558-11510	
Prüfinweis	Prüfservice: <a href="http://demo.a-sit.at/el_signatur/pdf-as/verify">http://demo.a-sit.at/el_signatur/pdf-as/verify</a>	

## Motivation

Text based PDF-signatures are widely used within Austrian E-Government (known as “Official Signatures”). This type of electronic signatures can be verified based on a paper printout as well. This is due to that text based PDF-signatures just sign the textual content of a given PDF document. For text extraction a canonical text representation is needed.

## Project description

- Goals
  - definition/specification of a canonical text extraction algorithm following reading direction
  - implementation of this algorithm to be used within PDF-AS Library (i.e. the core library used to create Official Signatures on PDF documents)
- Tasks
  - analysis of PDF document format (PDF/A1 [ISO 19005-1](http://www.iso.org/iso/19005-1))
  - definition of text extraction algorithm, key aspects are:
    - following reading direction
    - based on PDF/A specification
    - technical neutral
    - correct handling of footer and header
  - implementation as Java Library
    - to be integrated into the existing PDF-AS project
    - design of an appropriate interface to integrate arbitrary algorithms into PDF-AS

## Literature

- ISO 19005-1 (or its succeeding standard)
- E-Government Act (defining legal basis for Official Signatures)
- PDF-AS specification ([http://demo.egiz.gv.at/plain/projekte/signatur im e government/pdf signatur](http://demo.egiz.gv.at/plain/projekte/signatur_im_e_government/pdf_signatur))
- PDF-AS project (<http://pdf-as.egovlabs.gv.at/>)

## Deliverables

- specification of algorithm (in English)
- normative implementation of algorithm as Java Library to be used within PDF-AS
- master thesis

## Project schedule

- Duration 6 months

## Master Thesis

Studies:  INF  SEW  TEL

## Prerequisites

- Java and Java Security
- IT-Security
- XML-Digital Signatures

## Advisor / contact

[thomas.roessler@iaik.tugraz.at](mailto:thomas.roessler@iaik.tugraz.at)