Rechtsquellenstiftung des Schweizerischen Juristenvereins
Fondation des sources du droit de la Société suisse des juristes
Fondazione per le fonti giuridiche della Società svizzera dei giuristi

Mer enn bare TEI…

Bruk av tekstteknologier i SSRQ

Beni Ruef <bernhard.ruef@ssrq-sds-fds.ch>

Workshop
«Resultater og utfordringer i arbeidet med SSRQ»
UiB, 2. september 2022

Innhold

- (Kort) presentasjon av *Rechtsquellenstiftung des Schweizerischen Juristenvereins* («Den Sveitsiske Juristforeningens Rettskilderstiftelse»)
- Forskningsstand og samlingens karakteristikk
- Struktur av en SSRQ-edisjon
- Retrodigitalisering
- Lage registre automatisk
- SSRQs TEI-portal
- Trykte bøker i TEI-tidsalderen
- Håndtering av grafdata
- Viktigheten av metadata og hvordan håndtere dem
- Teknologier i bruk

Hvem/hva er den Sveitsiske Rettskilderstiftelse?

Forskningsinstitusjon, grunnlagt i 1898 av den sveitsiske juristforeningen som kommisjon, siden 1980 er den en uavhengig stiftelse.

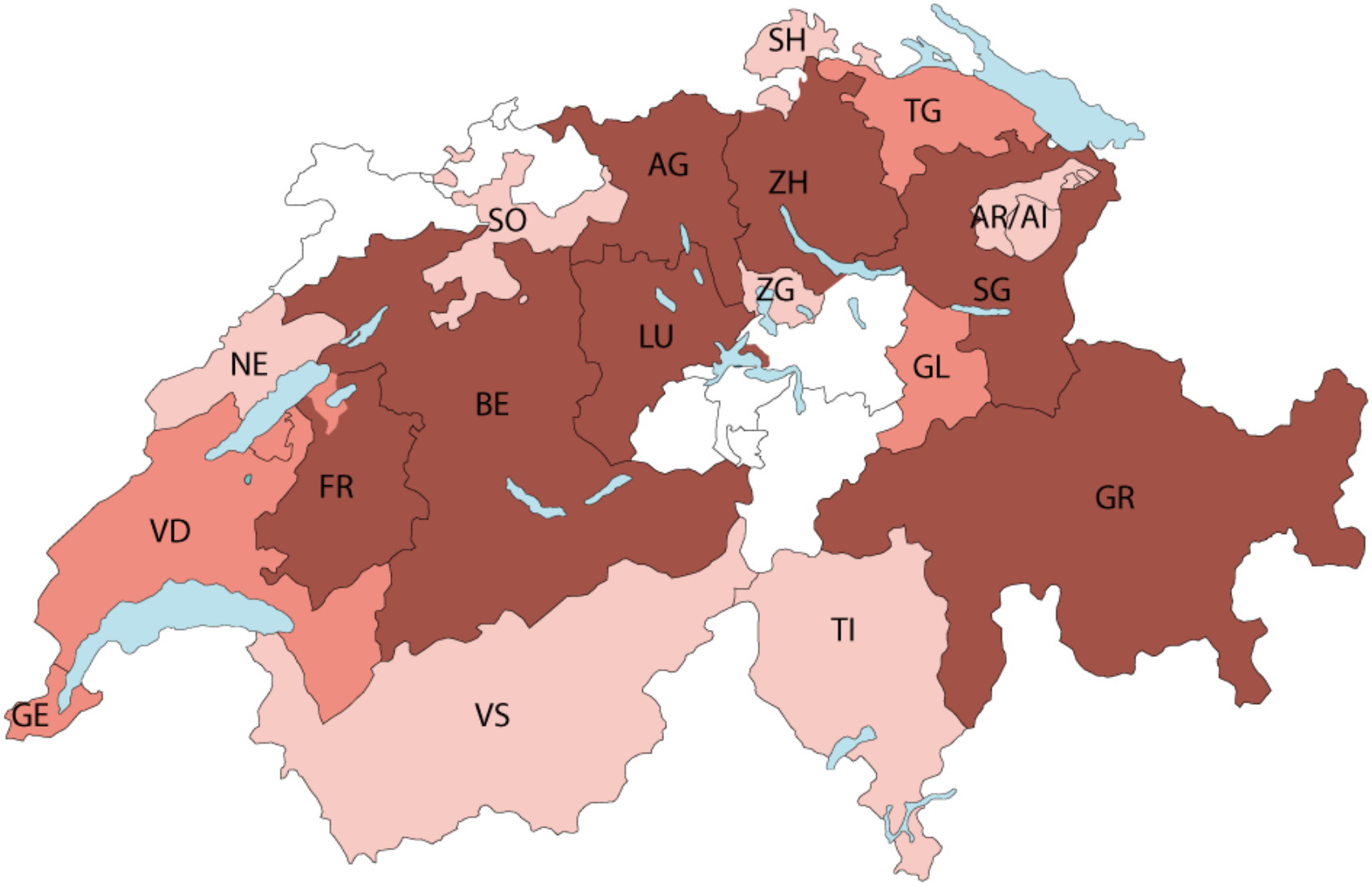
- Utgir såkalte *kritiske edisjoner* av sveitsiske rettskilder fra tidlig middelalder til 1798 i *Samlingen av sveitsiske rettskilder* (SSRQ: **S**ammlung **S**chweizerischer **R**echts**q**uellen).
- Tidlig middelalder betyr ca. fra 6. århundre.
- 1798 var året hvor Sveits ble invadert av franske styrker som betydde slutten av det «gamle edsforbundet» (*Alte Eidgenossenschaft*, *Ancien Régime*) og begynnelsen av den «Helvetiske Republikken» som var en fransk vasallstat.

Hva er rettskilder?

SSRQ tolker begrepet veldig bredt:

- konstitusjoner
- traktater
- bylover (*town charts*)
- rett til bruk (*rights of use*)
- kataloger av varer (*catalogues of goods*)
- dommer og andre dokumenter fra domstoler
- osv.

Forskningsstand (dekningen)



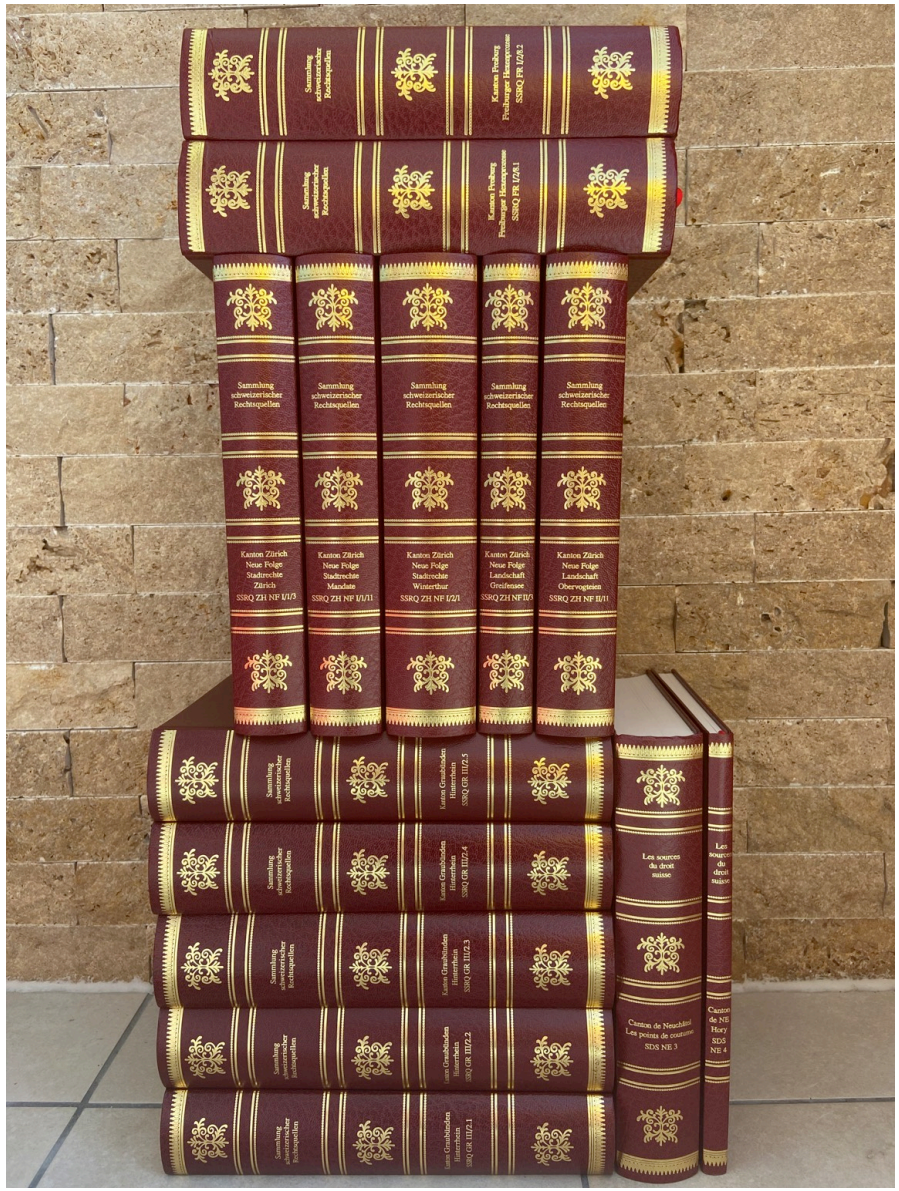
8 eller flere edisjoner | 4–7 edisjoner | 1–3 edisjoner

Noen tall…

- hittil 140 edisjoner (≠ fysiske bind) tilsvarende ca. 48’000 artikler (dokumenter) og ca. 93’000 trykksider
- 12 løpende prosjekt hvorav 8 er basert på XML/TEI
  - tysk: FR, GR, LU, SG, SH (2), TG, VS
  - fransk: GE, NE, VD
  - italiensk: TI

Produksjon 2021–2022

En ny rekord: ni prosjekt kunne avsluttes i løpet av et halvt år!



Samlingens karakteristikk (og utfordringer…)

- svært forskjellige teksttyper
- diakron
- ikke-normaliserte språkvarieteter
- mange spesialtegn (som medfører bruk av MUFI)
  - vokaler med aksent: a̍, ǎ, åͤ, e̊ͣ, o̊ͤ, o̊ͧ, oͮ, u̍, ū, ůͤ, uͦ osv. osv.
  - symbol for mynter: bl.a. (florin), (nederlandsk pund), (groschen), (schilling), (krone)
- dobbelt flerspråklig
  - kildetekst (tysk, fransk, italiensk, latinsk og retoromansk)
  - redaktørtekst (tysk, fransk og italiensk)
- med andre ord: et veldig heterogent korpus!

Fordeling over tid

Hvordan ble/er edisjonene trykt/laget?

Måten gjenspeiler den teknologiske utviklingen siden 1898, dvs. i snart 125 år.

- til 1960-tallet: blysats
- fra 1960-tallet til ca. 1995: fotosats
- fra ca. 1995 til 2010: *FrameMaker*
- fra 2010 til 2018: *InDesign*
- siden 2011: LaTeX (fortvilt pga. *InDesign* og besluttet å gjøre alt selv…)
- TEI siden 2010 (retrodigitalisering) / 2012 (*digitally born*)

Anatomi av et SSRQ-bind

- fortegnelse av kildetekster
- innføring
- diverse lister (liste av arkiver, bibliografi, liste av forkortelser osv.)
- kildetekster
  - tittel
  - dato
  - innledende bemerkninger (f.eks. historisk ramme)
  - transkripsjon
  - manuskriptbeskrivelse (arkiv, materiale, format, segl osv.)
  - anmerkninger
  - flere noter (f.eks. relaterte tekster eller ettervirkninger)
- registre
  - personer, familier og organisasjoner
  - stedsnavn
  - lemma
  - nøkkelord

Retrodigitalisering (*SSRQ online*)

- Fase 1 (2008–2011)
  - alle sider skannet (G4 komprimert TIFF, dvs. bitonal, 600 dpi) og fiffet opp (*wiping margins* og *deskewing*)
  - kataloger av kildetekster OCR-behandlet og manuelt korrigert som muliggjør søking i dokumentenes titler
  - utvikling av en *viewer* (bildebetrakter) for faksimiler
- Komplett OCR (2017)
  - analoge bind (1898–1996) OCR-behandlet
    - brukt ABBYY FineReader 14
    - trent programmet avhengig av språk og periode (typografien forandret seg over tid)
    - korrekt identifikasjon av ſ (LATIN SMALL LETTER LONG S) og tegn som ů
    - resultat gjort tilgjengelig som PDF
  - post-1995 («digitale») bind også gjort tilgjengelig som PDF

Lage registre automatisk

- Problem
  - Våre registre inneholder informasjon om deres oppslagsord, ikke bare sidenumre.
  - I følge tradisjonen oppgir de også linjenummer.
  - Tidligere ble registre laget manuelt (sic!), selv med *FrameMaker* og *InDesign*…
- Løsning
  - LaTeX-makroer (`\persname` etc.) tagger oppslagsord og lager lister av forekomster.
  - Informasjon om oppslagsord er lagret i en database.
  - Et Perl-skript lager LaTeX-koden for registret fra forekomstene og databasen.

Hva er så spesielt med SSRQs TEI-portal?

- fullstendig basert på *TEI Processing Model*
- kompleks kritisk noteapparat med meget nøstete TEI-elementer, f.eks.
  `<app><lem/><rdg><unclear/></rdg></app>`
- fyldige semantiske annotasjoner (`<date>`, `<measure>`, `<persName>`, `<placeName>` osv.)
- komplett TEI-skjema og validering til og med attributverdier dokumentert i ODD
- flerspråklig: navn (*labels*) og attributverdier oversatt til tysk, fransk og engelsk (og veldig snart også italiensk)

Hvorfor/hvordan lage en trykt bok i TEI-tidsalderen

- Ja, en trykt bok er fremdeles nødvendig!
  - som uforanderlig publikasjon egnet for referanse
  - som langtidslagring *(ikke le)*
  - for prosjektets festlige avslutning *(ikke le her heller)*

Krav og valg

- Krav
  - veldefinert, klar struktur og rent oppsett
  - støtter typografisk kompleksitet (registre, orddeling, linjenumre osv.) og kvalitet (ligaturer, *kerning* osv.)
- Valg
  - XSL-FO
    - utilstrekkelig til formatering av et kompleks kritisk noteapparat
    - typografisk kvalitet tilfredsstiller ikke høye krav med mindre det kjøpes et kommersielt produkt
  - LaTeX
    - har en lang tradisjon for å lage kritiske edisjoner (med pakker som f.eks. *reledmac*)
    - LaTeX-koden kan bevare TEIs semantikk: leselig for mennesker, enklere å feilsøke
    - SSRQ har LaTeX-kompetanse fra før

Hvordan omforme XML/TEI til LaTeX

- tradisjonelt gjort med XSLT
- fordeler ved bruk av TEI ODD og *Processing Model*
  - mange TEI-elementer bruker same `<model>` for både web og utskrift som resulterer i mye mindre kode
  - veldig rask implementering (oversettes direkte til XQuery-funksjoner)

Håndtering av grafdata

- I dag er oppslagsord lagret i tre forskjellige databaser basert på to forskjellige teknologier: XML (*eXist-db*) og RDF (*Apache Jena Fuseki*).
- Men disse oppslagsord ($\equiv$ entiteter) er svært sammenkoblede.
- Med andre ord: de kan representeres som *graf*.
- Hvordan lagrer man grafdata?
- Vår mål: kombinere elegansen av en grafdatabase med stabiliteten av en relasjonal database
- Vår fremgangsmåte: lage tabeller og SQL-kode automatisk fra ontologien som beskriver grafen (et slags ORM)

Viktigheten av metadata og hvordan håndtere dem

- Metadata er helt essentiell!
- Det finnes ingen adgang til data i digital humaniora uten metadata; fulltekst-søk er aldri noen erstatning.
- Mange typer og nivåer av metadata
  - om kildetekst
  - om transkripsjon
  - registerdata
  - metadata om registerdata (f.eks. når ble person X født?), dvs. metadata om metadata
  - metadata om metadata om registerdata (referanse for fødselsdato av person X), dvs. metadata^3 :-)
- Hvor/hvordan skal man lagre metadata?
- Hvordan skal man gjøre dem tiljengelig?

Teknologier i bruk

(ufullstendig liste…)

- HTML5
- Perl
- LaTeX (egentlig XeLaTeX og litt TeX)
- XML, XSLT, XQuery, Schematron
- Python, RDF, SPARQL
- TEI ODD, TEI Processing Model
- PostgreSQL, SQL
- OCaml
- og til sist men ikke minst: alt er bundet sammen med Unix shell!