Aalto University
School of Science

UNIVERSITY OF HELSINKI

SKS Suomalaisen Kirjallisuuden Seura

SeCo

HELDIG
Helsinki Centre for Digital Humanities

HELSINKI INSTITUTE FOR SOCIAL SCIENCES AND HUMANITIES

Aineistojen muuntaminen ja jalostaminen linkitetyksi avoimeksi dataksi: 1,2 miljoona kirjettä ja 100 000 toimijaa

Suomalaisen Kirjallisuuden Seura
27.5.2025

Petri Leskinen, Aalto-yliopisto
Senka Drobac, Helsingin yliopisto

Sisältö

- Lähdemateriaali
- Datamuunnos
- Datan rikastaminen
- Tietomalli

Aalto University
School of Science

UNIVERSITY OF HELSINKI

SKS Suomalaisen Kirjallisuuden Seura

SeCo

HELDIG
Helsinki Centre for Digital Humanities

HELSINKI INSTITUTE FOR
SOCIAL SCIENCES
AND HUMANITIES

Lähdemateriaali

Lähdemateriaali

- **Lukuisilta eri osapuolilta**
  - Datan määrä vaihtelee, 25 - 360.000 kirjettä
  - Ajallinen rajaus vuoteen 1917
- **Useissa eri formaateissa**
  - Word, Excel, CSV, JSON, RDF
- **Tiedon kattavuus vaihtelee**
  - Aika, kirjeiden määrä, lähetys- ja vastaanottopaikat
  - Henkilöiden biografiset yksityiskohdat

Aalto University
School of Science

UNIVERSITY OF HELSINKI

SKS Suomalaisen Kirjallisuuden Seura

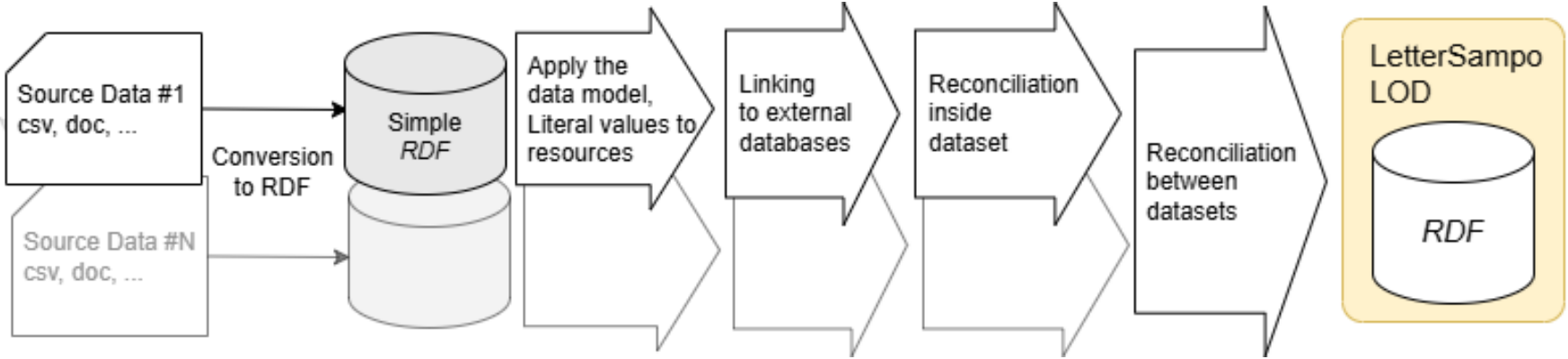
SeCo

HELDIG
Helsinki Centre for Digital Humanities

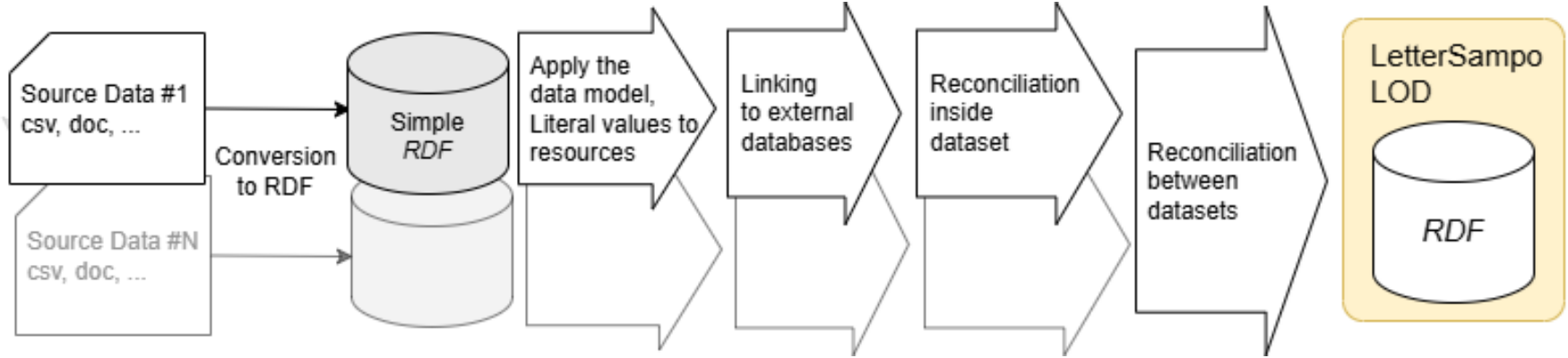
HELSINKI INSTITUTE FOR
SOCIAL SCIENCES
AND HUMANITIES

Datamuunnos

Process pipeline

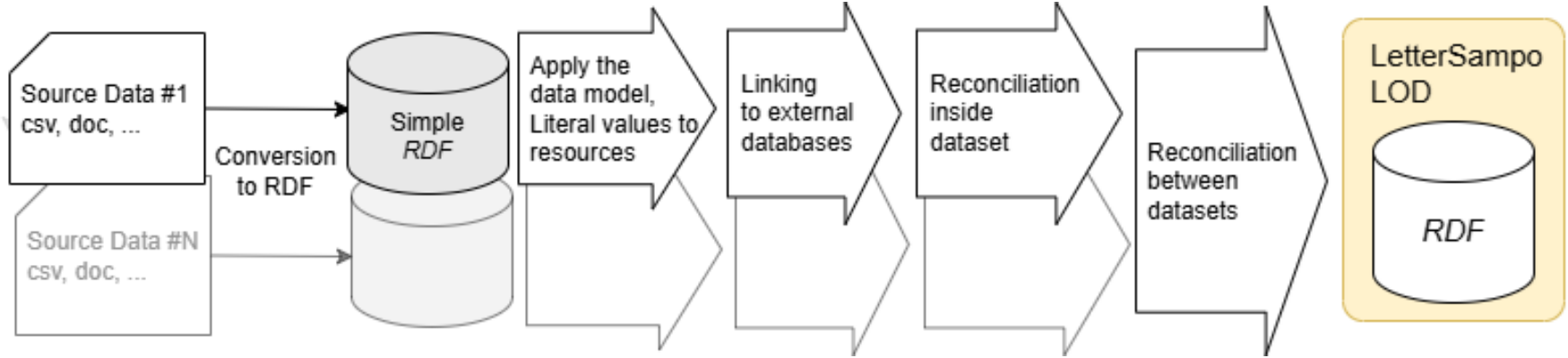


Process pipeline



- **Data transformation**
- Harmonization
- Deduplication
- Data Linkage

Process pipeline



- **Data transformation**
  - **From various data formats:**
    Word, Excel, CSV, JSON, RDF
- Harmonization
- Deduplication
- Data Linkage

Word files

- **Different document structure**
  - Categories not in the same order, different titles
  - Data for several people in one document
- **Inconsistent formatting**
  - Mixing tabs/spaces
  - Text in multi lines
  - Other correspondences
  - Numerous small inconsistencies

Multi lines

Number ↓ of letters

Suomen Eläinsuojeluyhdistys	1929	1
Suomen Evankelinen Seura	1918	1
Suomen ev.lut. Pyhäkouluyhdistys	1922	1
Suomen Historiallinen Seura	1912-1925	13
Suomen kenraalikuvernööri (:n toimisto), /	1900-1916	30
# liitteinä suomennoksia	~~1900-1916~~	~~30~~
Suomen kielen sanakirjaosakeyhtiö	1916	2
Suomen Kirkkohistoriallinen Seura	1902	1
Suomen Kirkollisviraston leski- ja orpokassa	1900-1919	2
Suomen Kirkon Pappisliitto	1918-1926	6
Suomen Kirkon Seurakuntatyön Keskusliitto (liite)	1918-1934	23
Suomen Kirkon Sisälähetysseura (liite) /	1917	
##(*ks. myös Aarnisalo, Otto ja Suomen Piplia-seurat*)	~~1917~~	~~1~~
Suomen Kotiseutututkimuksen keskusvaliokunta	1917	1

Multi lines

Detschy, Serafine	s.a.	1
Dillon, E.J.D. liitt. Dillon, E.J.D. → Mr Dillon s.a.,1 kpl	1891, s.a.	6
Dingeldey, Ludvig	s.a.	3
Dmitrieff, Heléne	1906, s.a.	4
Donner, Otto (kts. myös Donner, Minette)	1885-1904, s.a.	11
Donner (o.s. Munck), Wilhelmina Sofia Ch. (Minette) & Donner, O.	1899	1
Dühren, C.J. von	1904	1

Donner (o.s. Munck) ; Wilhelmina Sofia /
Ch. (Minette) ; Donner, O.

Gallen-Kallela Museum / FINNA

Results of a database query using Finna API Service

Albert Edelfelts brev API


```
{
        "url": "http://edelfelt.sls.fi/api/letters/1/",
        "id": 1,
        "title": "Kiala den 16 Juni 67.",
        "events": [ "http://edelfelt.sls.fi/api/events/1983/", ... ],
        "web_url": "http://edelfelt.sls.fi/brev/1/kiala-den-16-juni-67/",
        "date": "1867-06-16",
        "urn": "URN:NBN:fi:sls-537-1403107521400",
        "locations": [ "http://edelfelt.sls.fi/api/locations/273/" ]
}
...
{
        "url": "http://edelfelt.sls.fi/api/events/1995/",
        "id": 1995,
        "title": "Alexandra Edelfelt och Carl Albert Edelfelt har varit på resa; de har besökt Åbo och Stockholm.",
        "letter": "http://edelfelt.sls.fi/api/letters/2/",
        "persons": [ "http://edelfelt.sls.fi/api/persons/466/",
                "http://edelfelt.sls.fi/api/persons/786/"
        ],
        "mentioned_locations": [ "http://edelfelt.sls.fi/api/locations/71/",
                "http://edelfelt.sls.fi/api/locations/78/"
        ]
}
```

Simple RDF

```
records:NL__00009996 a :MetadataRecord ;
    sender "Borg, Einar",
        "Heikel, O. J.",
        "Järventaus, H. A.",
        "Kauppinen, K.",
        "ym." ;
    :sender_as_recorded "Borg, Einar ; Heikel, O. J. ; Järventaus, H.A. ; Kauppinen, K. ;  ym." ;
    :recipient "Juho Rudolf Koskimies (Forsman)" ;
    :recipient_as_recorded "Juho Rudolf Koskimies (Forsman)" ;
    :sending_date "1916" ;
    :number_of_letters "5" ;
    :archival_organization_as_recorded "National Library of Finland" ;
    :fonds_as_recorded "Coll. 108 Koskimies, Juho R." ;
    :source_filename "COLL108.docx" ;
    :note "# KK" ;
    :archival_location "COLL.108.27" ;
    dct:source <http://ldf.fi/coco/source/nationallibrary> .
```

← Actors

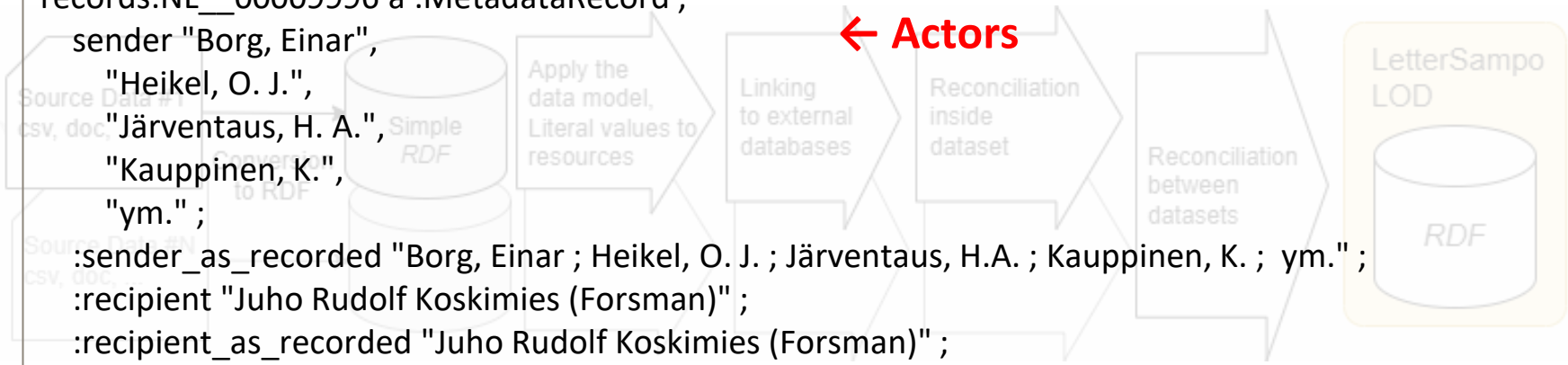
← Timespan

← Letters

← Contributor

← Fonds

← Additional archival metadata



Aalto University
School of Science

UNIVERSITY OF HELSINKI

SKS Suomalaisen Kirjallisuuden Seura

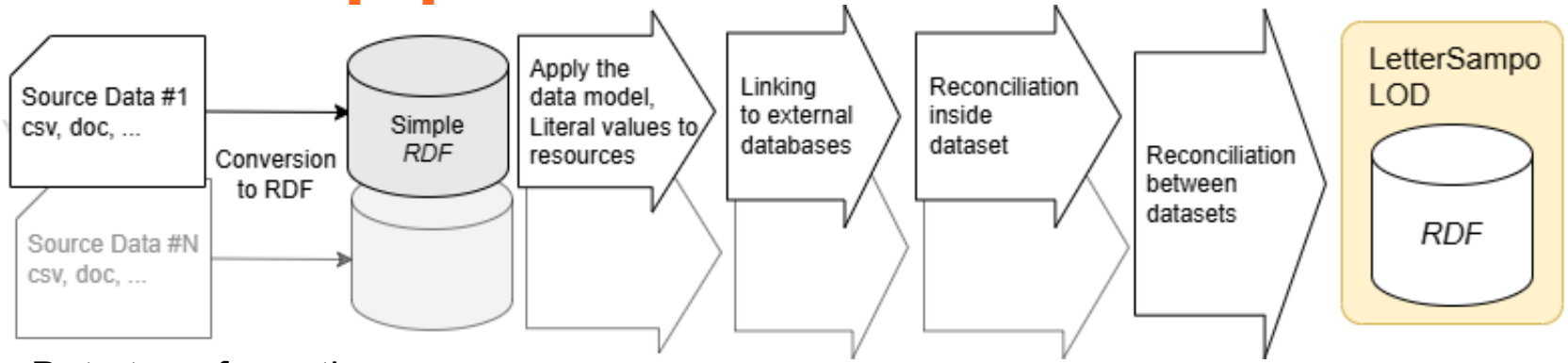
SeCo

HELDIG
Helsinki Centre for Digital Humanities

HELSINKI INSTITUTE FOR
SOCIAL SCIENCES
AND HUMANITIES

Datan rikastaminen

Process pipeline

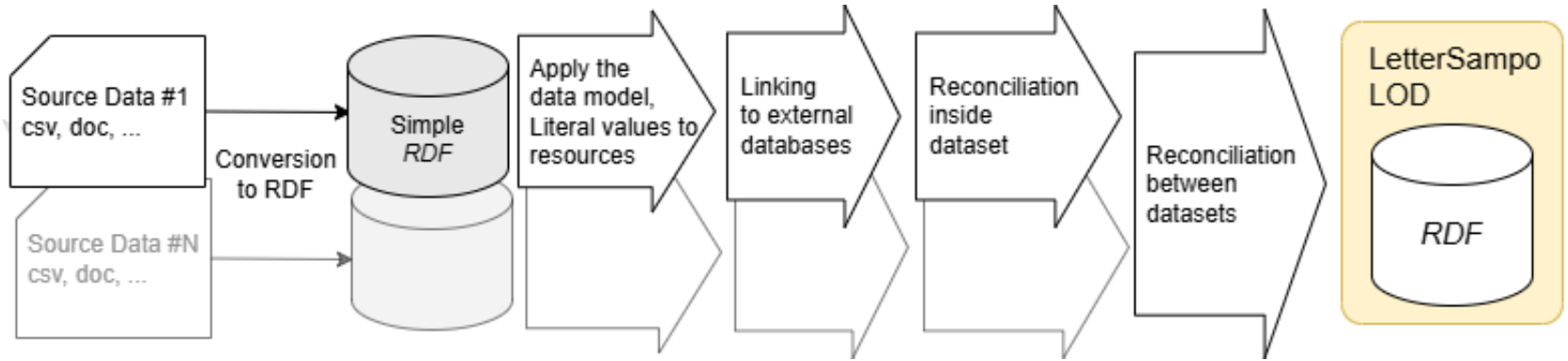


- Data transformation
- **Harmonization**
  - **Into one uniform data schema**
- Deduplication
  - Merge same actor mentioned in different datasets
- Data Linkage
  - To external data sources
  - Enrich with e.g. biographical details and family relations

Esimerkkejä

- **Henkilönimet**
  - "Haartman, Axel o. Hedvig f. Stolpe"
    - *Haartman, Axel*
    - *Haartman (Stople), Hedvig*
- **Elin- tai toiminta-aika**
  - Forsström, Petter (1877-1967)
  - Kekoni, Agnes *(fl. 1876-1944)*
  - Förlags Ab Söderström & Co *(fl. 1800-1949)*
- **Arkistokokoelmat**
  - “Hilja Haahden arkisto”
    - *Haahti, Hilja*
  - “SLSA 1319 Släkten Munsterhjelms arkiv”
    - *Munsterhjelm, Hjalmar; Munsterhjelm, Anders Lorentz; Munsterhjelm, Anders Gustaf ; Munsterhjelm, Sofie Johanna; … (linked to Finna)*

Process pipeline

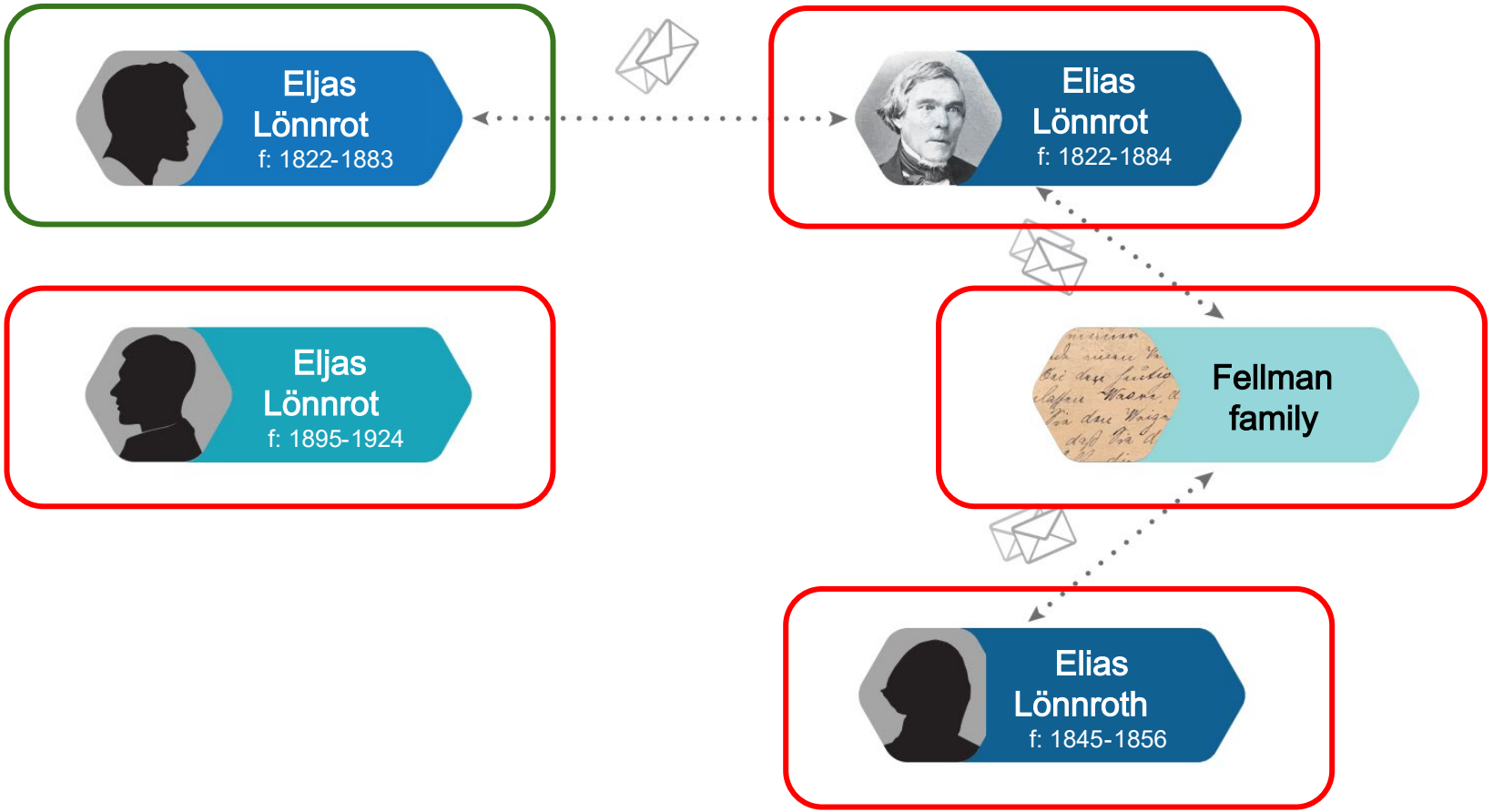


- Data transformation
- Harmonization
- **Deduplication**
  - **Merge same actor mentioned in different datasets**
- Data Linkage
  - To external data sources
  - Enrich with e.g. biographical details and family relations

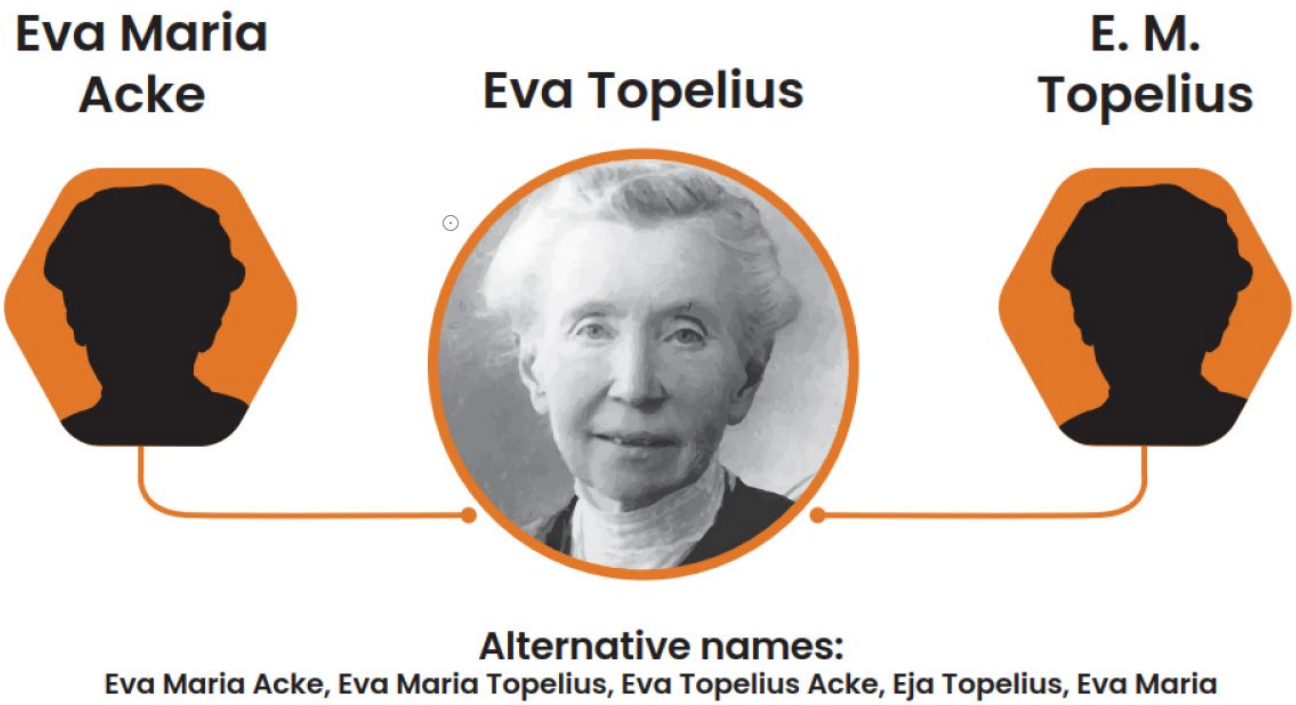
Toimija- ja paikkaontologiat

- **Henkilö, Organisaatio, Perhe/Suku, tuntematon**
  - Ulkopuoleiset tietolähteet
    - *SampoSampo*
    - *Biografiset tiedot, perhe- ja ihmissuhteet*
- **Paikat**
  - Ulkopuoleiset tietolähteet
    - *Koordinaatit, hiearkia*

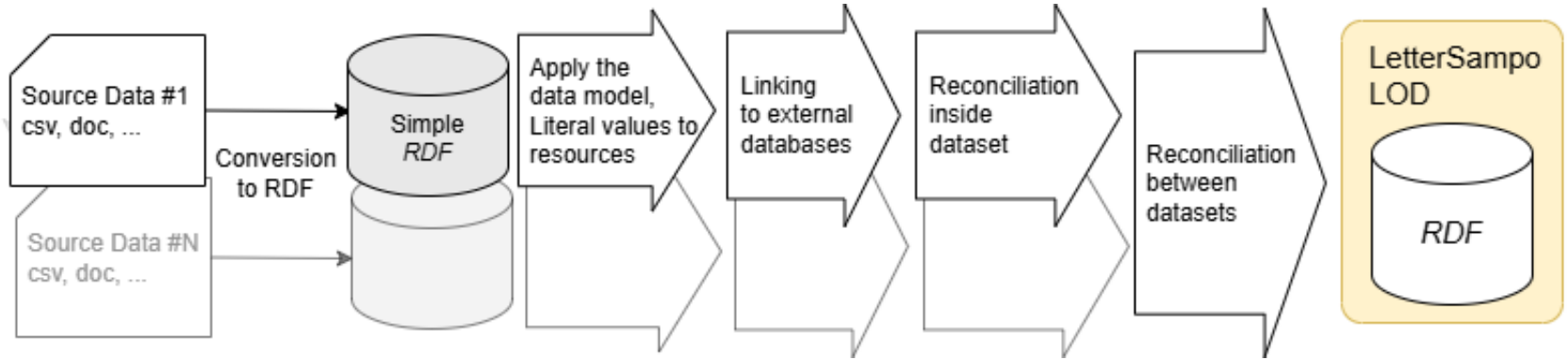
Actor deduplication



Actor deduplication



Process pipeline



- Data transformation
- Harmonization
- Deduplication
- **Data Linkage**
  - To external data sources
  - Enrich with e.g. biographical details and family relations

Vocabulary of Finnish Actors

- SampoSampo

  Historical Finnish People, Organizations, and Places
  Biographical information about 230000 historical people related to Finland
  Name variations, gender, years and places of birth and death ...

- Collected from multiple datasources:

1. AcademySampo (23000 entries)
2. BiographySampo (22500)
3. Wikidata (84000)
4. Kanto (29000)
5. BookSampo (9800)
6. ParliamentSampo (2100)
7. Norssi Highschool Alumni (2600)
8. Union List of Artist Names (1500)
9. Wikitree (12000)
10. Geneanet (144000)

    Parts of CoCo data providing biographical details

11. Edelfelt Letters (3300)
12. Snellman Letters (3300)
13. Åbo Akademi University Library letters (1570)

Vocabulary of Finnish Actors

Eva Topelius (Q4455334)

Finnish painter

Eva Maria Topelius | E. M. Topelius | E. Topelius | Eva Maria Acke

In more languages

Configure

Language	Label	Description	Also known as
English	Eva Topelius	Finnish painter	Eva Maria Topelius E. M. Topelius E. Topelius Eva Maria Acke
Finnish	Eva Topelius	suomalainen taidemaalari	E. Topelius E. M. Topelius Eva Maria Topelius Eva Maria Andersson Eva Maria Acke Eva Topelius Acke Eva Topelius-Acke
Swedish	Eva Topelius	finlandssvensk målare	E. Topelius E. M. Topelius Eva Maria Topelius Eva Maria Acke

KÄYTETTÄVÄ NIMENMUOTO	Topelius, Eva, 1855-1929
TYYPPI	henkilö
HENKILÖN TUNNISTE	Asteri ID: 000122880
SYNTYMÄAIKA	1855
KUOLINAIKA	1929
SYNTYMÄPAIKKA	Uusikaarlepyy (yso-paikat)
KUOLINPAIKKA	Vaxholm, Ruotsi
HENKILÖN AMMATTI TAI TEHTÄVÄ	taidemaalari (Metatietosanasto)
HENKILÖN TOIMINTA-ALA	maalaustaide (YSO)
HENKILÖÖN LIITTYVÄ HENKILÖ	Topelius-Acke, Eva, 1855-1929

♀ Eva Maria Topelius

- Born in 1855 - Nykarleby
- Deceased March 23, 1929 - Vaxholm, Sverige, aged 74 years old

1 file available

Parents

- Zacharias Topelius, *Verkligt statsråd* 1818-1898
- Maria Emilie Lindqvist 1821-1885

Eva Topelius-Andersson

★ Uusikaarlepyy 4.9.1855 † Vaxholm 1929

taidemaalari

Lähteet: Kansallisbiografia, Kansallisbiografian sisällysluettelo

ISNI | VIAF | Wikipedia | Wikidata

Challenges

- General issues of record linkage, e.g., namesakes, name variations
- Errors in source data
  - Typos
  - Uncertainty in letter dates
  - Erroneous dates
- No (machine readable) data available about some actors

Record Linkage, Current Implementation

- Linkage to external databases
- Deduplication between datasets based on external matches

- Data Enrichment from external sources
  - Biographical details
  - Occupations
  - Personal relations
  - Images

Record Linkage, Current Implementation

- Using Python package SPLink:
  - Unsupervised learning
  - Term Frequency learning
  - Alternative Python packages: Dedupe, Record Linkage
- Similarity based on:
  - Three different name variations considering
    - *Family Names*
    - *Given Names*
    - *Middle Names*
    - *Initials*
  - Quantiles of letter activity period
    - *Compared to lifespans of external actors*
  - Existing linkage
    - *Cases of Snellman and Edelfelt*

Aalto University
School of Science

UNIVERSITY OF HELSINKI

SKS Suomalaisen Kirjallisuuden Seura

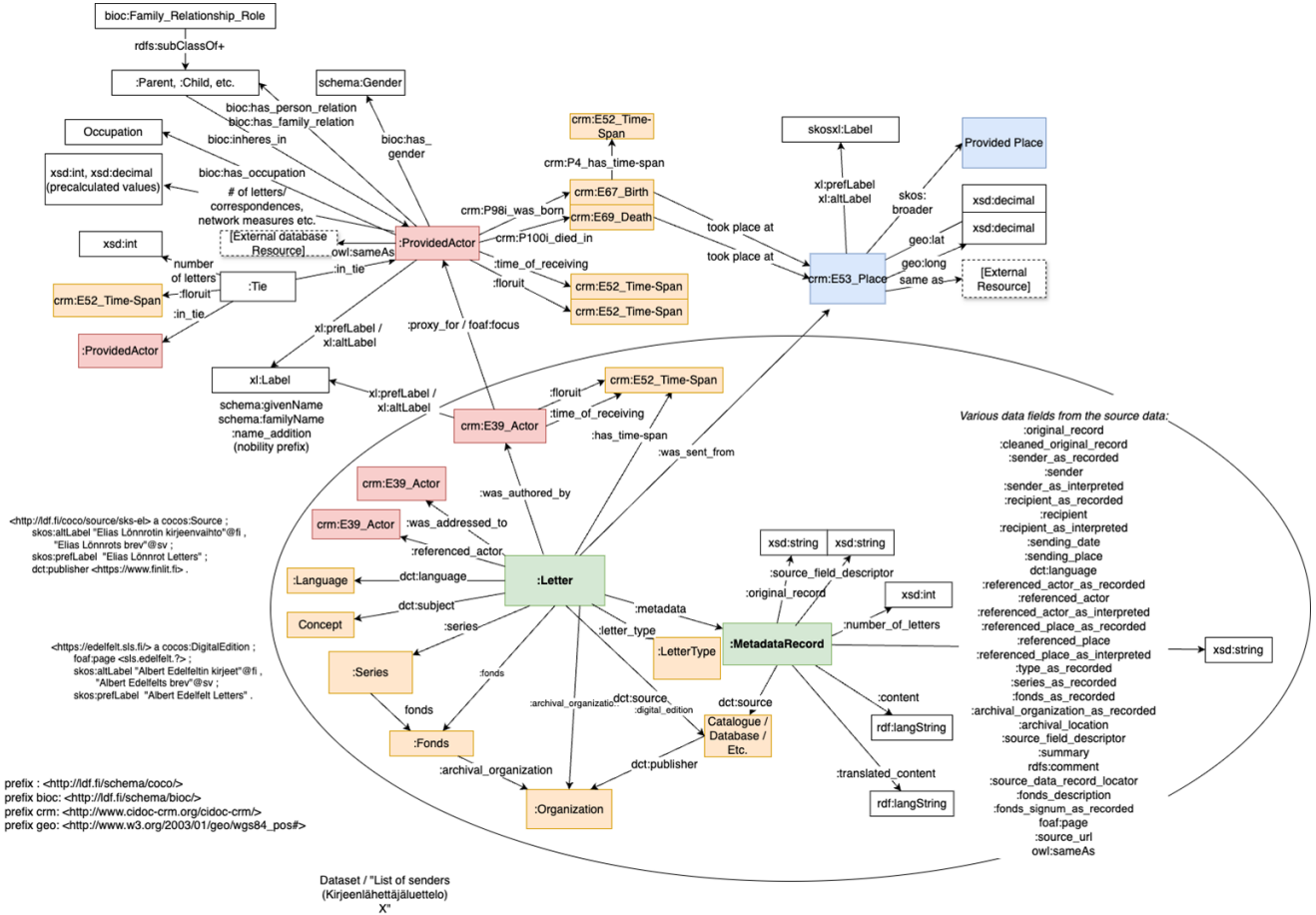
SeCo

HELDIG
Helsinki Centre for Digital Humanities

HELSINKI INSTITUTE FOR
SOCIAL SCIENCES
AND HUMANITIES

Tietomalli

Tietomalli

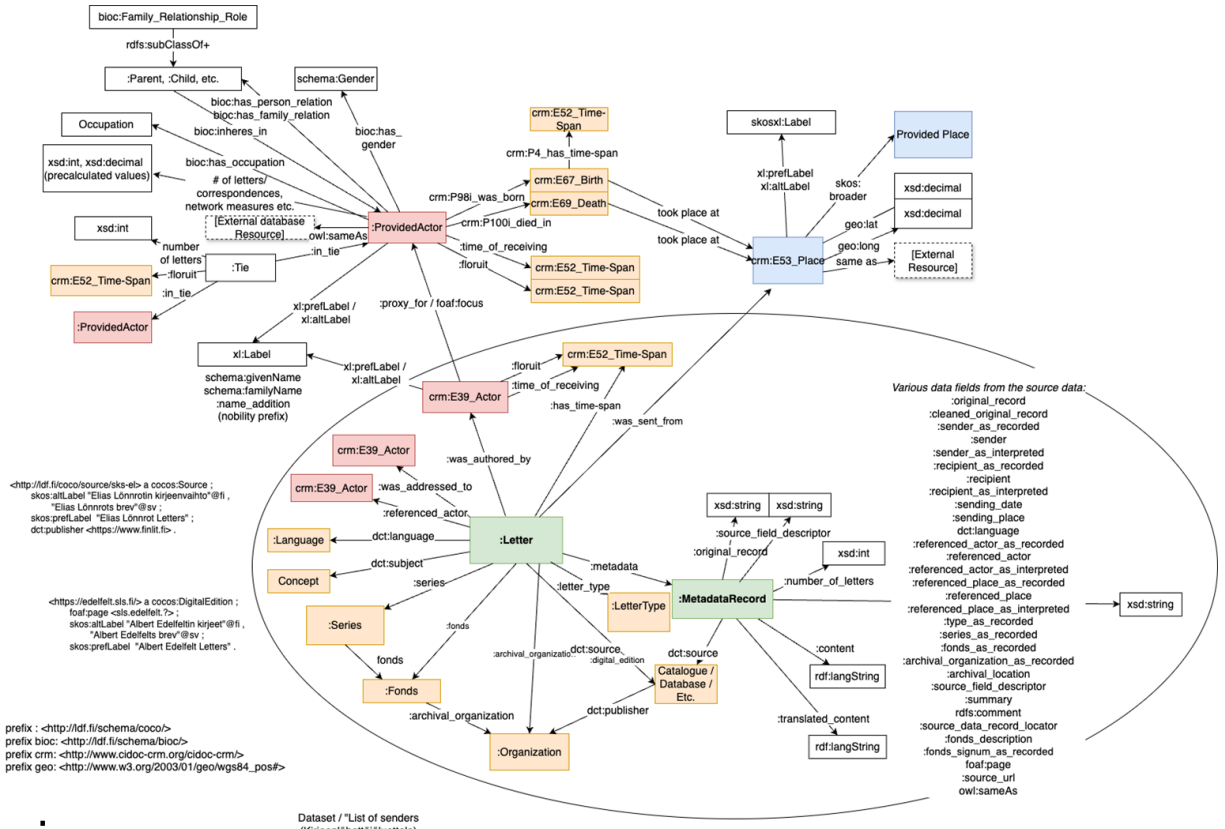

bioc:Family_Relationship_Role
rdfs:subClassOf+
:Parent, :Child, etc.
schema:Gender
bioc:has_person_relation
bioc:has_family_relation
bioc:inheres_in
bioc:has_gender
Occupation
xsd:int, xsd:decimal (precalculated values)
bioc:has_occupation
# of letters/ correspondences, network measures etc.
xsd:int
[External database Resource]
owl:sameAs
:ProvidedActor
number of letters
:in_tie
:Tie
:floruit
crm:E52_Time-Span
:in_tie
:ProvidedActor
crm:E52_Time-Span
crm:P4_has_time-span
crm:E67_Birth
crm:E69_Death
crm:P98i_was_born
crm:P100i_died_in
:time_of_receiving
:floruit
crm:E52_Time-Span
crm:E52_Time-Span
skosxl:Label
xl:prefLabel
xl:altLabel
Provided Place
skos: broader
xsd:decimal
xsd:decimal
geo:lat
geo:long
same as
[External Resource]
took place at
took place at
crm:E53_Place
xl:prefLabel / xl:altLabel
:proxy_for / foaf:focus
xl:Label
schema:givenName
schema:familyName
:name_addition
(nobility prefix)
xl:prefLabel / xl:altLabel
crm:E39_Actor
:floruit
crm:E52_Time-Span
:time_of_receiving
:has_time-span
:was_sent_from
crm:E39_Actor
:was_authored_by
crm:E39_Actor
:was_addressed_to
:referenced_actor
:Letter
:Language
dct:language
Concept
dct:subject
:Series
:series
:fonds
:Fonds
fonds
:archival_organization
:Organization
:archival_organization
dct:source
:digital_edition
dct:source
Catalogue / Database / Etc.
dct:publisher
:metadata
:letter_type
:LetterType
:MetadataRecord
xsd:string
xsd:string
:source_field_descriptor
:original_record
xsd:int
:number_of_letters
:content
rdf:langString
:translated_content
rdf:langString
Various data fields from the source data:
:original_record
:cleaned_original_record
:sender_as_recorded
:sender
:sender_as_interpreted
:recipient_as_recorded
:recipient
:recipient_as_interpreted
:sending_date
:sending_place
dct:language
:referenced_actor_as_recorded
:referenced_actor
:referenced_actor_as_interpreted
:referenced_place_as_recorded
:referenced_place
:referenced_place_as_interpreted
:type_as_recorded
:series_as_recorded
:fonds_as_recorded
:archival_organization_as_recorded
:archival_location
:source_field_descriptor
:summary
rdfs:comment
:source_data_record_locator
:fonds_description
:fonds_signum_as_recorded
foaf:page
:source_url
owl:sameAs
xsd:string
<http://ldf.fi/coco/source/sks-el> a cocos:Source ;
skos:altLabel "Elias Lönnrotin kirjeenvaihto"@fi ,
"Elias Lönnrots brev"@sv ;
skos:prefLabel "Elias Lönnrot Letters" ;
dct:publisher <https://www.finlit.fi> .
<https://edelfelt.sls.fi/> a cocos:DigitalEdition ;
foaf:page <sls.edelfelt.?> ;
skos:altLabel "Albert Edelfeltin kirjeet"@fi ,
"Albert Edelfelts brev"@sv ;
skos:prefLabel "Albert Edelfelt Letters" .
prefix : <http://ldf.fi/schema/coco/>
prefix bioc: <http://ldf.fi/schema/bioc/>
prefix crm: <http://www.cidoc-crm.org/cidoc-crm/>
prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
Dataset / "List of senders (Kirjeenlähettäjäluettelo) X"

Tietomalli

- Metadata Records

- Letters
- Actors
- Fonds
- Places
- Contributors

- Occupations
- Lettertypes
- Times
- … supporting ontologies



Tietomalli

letters:NL__00009996_0000001 a :Letter ;
:estimated_year 1916 ;
:fonds fonds:f2169107577047552292 ;
:has_time-span <http://ldf.fi/coco/times/time_1916-01-01T00:00:00-1916-12-31T23:59:59-1916-01-01T00:00:00-1916-12-31T23:59:59> ;
:metadata records:NL__00009996 ;
:original_data_provider sources:nationallibrary ;
:type :missing_value ;
:was_addressed_to actors:p2733526637996731453 ;
:was_authored_by actors:p16414152980933005 89,
actors:p16872944935414520 16,
actors:p18862810864766164 65,
actors:p40821138263642680 26,
actors:p59716694130309147 71 ;
dct:source sources:nationallibrary ;
skos:prefLabel "1916: Borg, Einar; Heikel, O. J.; Järventaus, H. A.; Kauppinen, K.; ym. -> Juho Rudolf Koskimies (Forsman) [NL__00009996]" .

Aalto University
School of Science

UNIVERSITY OF HELSINKI

SKS Suomalaisen Kirjallisuuden Seura

SeCo

HELDIG
Helsinki Centre for Digital Humanities

HELSINKI INSTITUTE FOR
SOCIAL SCIENCES
AND HUMANITIES

Kiitos!