Clustering, Extracting and Linking Bibliographic Work Entities

Andreas Andersson

National Library of Sweden

KBR Royal Library of Belgium

19/09/2023

Agenda

- Works as anonymous entities 2018 - 2023
- Works as linked entities 2024 -
  - Title clustering
  - Selection
  - Normalisation
  - Extraction
- Challenges

National Library of Sweden

BIBFRAME in Libris

KB Base Vocabulary (KBV)

Ontology	Properties	Classes	Example
BIBFRAME	213	204	bf:code
schema.org	42	18	sdo:birthDate
BFLC	12	9	bflc:encodingLevel
DC	21	7	dcterms:replaces
SKOS	21	3	skos:prefLabel
FOAF	6	3	foaf:name
BIBO	3	3	bibo:map
PROV	4	2	prov:entity
KBV unique	121	153	kbv:descriptionUpgrader
"marc"	549	233	marc:bib880-1

National Library of Sweden

Work entities 2018 - 2023

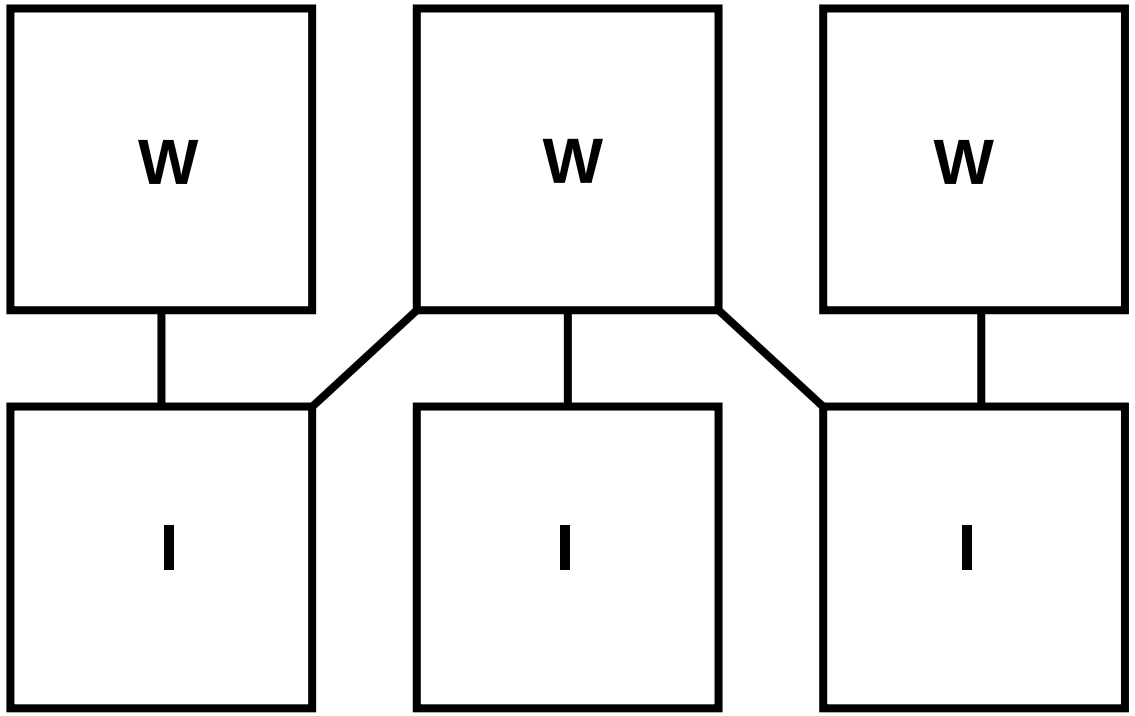
- From MARC record to RDF graph
- BIBFRAME conceptual model
- Identify abstract Work level

Instance	
Identifier	bf:identifiedBy
Extent	bf:extent
Summary	bf:summary
Publication	kbv:publication
Work	bf:instanceOf
Primary contributor	bf:contribution bflc:PrimaryContributor
Content type	bf:content
Subject headings	bf:subject
Classification	bf:classification


```
"@id": "https://libris.kb.se/p719rbl11qn9b2m#it",
"@type": "Print",
"extent": [
  {
    "@type": "Extent",
    "label": [
      "239 s."
    ]
  }
],
"identifiedBy": [
  {
    "@type": "ISBN",
    "value": "9789174293814"
  }
],
"instanceOf": {
  "@type": "Text",
  "language": [
    {
      "@id": "https://id.kb.se/language/swe"
    }
  ],
  "genreForm": [
    {
      "@id": "https://id.kb.se/marc/Novel"
    }
  ],
  "contentType": [
    {
      "@id": "https://id.kb.se/term/rda/Text"
    }
  ],
  "classification": [
    {
      "code": "813.54",
      "@type": "ClassificationDdc",
      "edition": "full",
      "editionEnumeration": "23/swe"
    }
  ],
"issuanceType": "Monograph",
"hasDimensions": {
  "@type": "Dimensions",
  "label": [
    "22 cm"
```


National Library of Sweden

Stand-alone Work entities



National Library of Sweden

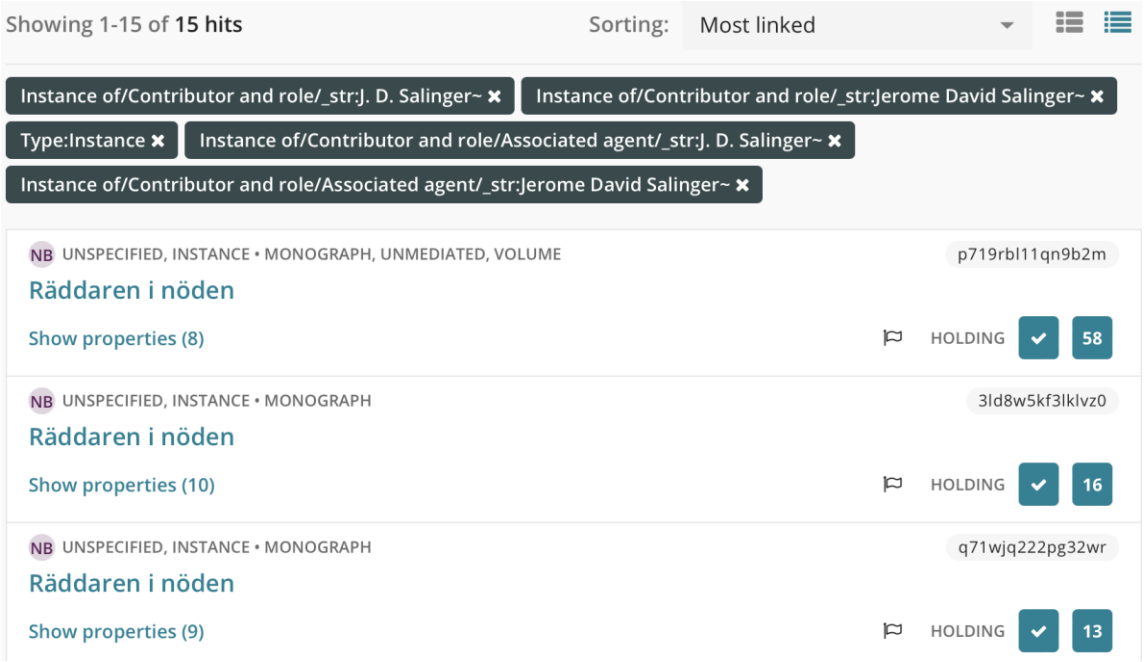
Exctracting Work entities

1. Title clustering
2. Selection
3. Normalisation
4. Extraction

Step 1: Title clustering

A. Search query (title + contributors)

- kbv:hasTitle.bf:mainTitle: Räddaren i nöden AND bf:contribution:[J. D. Salinger OR Jerome David Salinger]
- 15 million instances → 1,6 million clusters

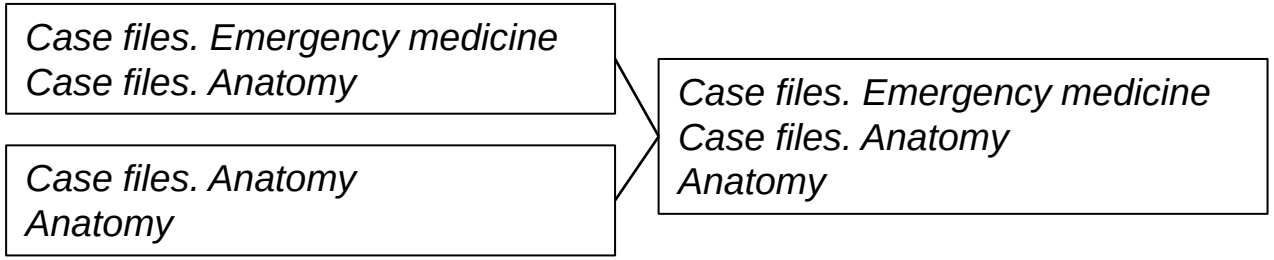



```
[q:[*], @type:[Instance],
hasTitle.mainTitle:[Räddaren i
nöden], or-
instanceOf.contribution._str:[J.
D. Salinger, Jerome David
Salinger], or-
instanceOf.contribution.agent._str
:[J. D. Salinger, Jerome David
Salinger]]
```

National Library of Sweden

Step 1: Title clustering

B. Merging similar clusters



- 1,6 million cluster → 1,4 million clusters

C. Compare all titles

- bf:mainTitle + bf:subtitle + bf:hasPart + bf:partNumber + bf:partName, marc:parallelTitle + marc:equalTitle - punctuation
- Ignore generic subtitles (*a comedy*, *eine Erzählung*) and main titles (*Works*, *Selections*)
- 1,4 million clusters → 1 million clusters

National Library of Sweden

Step 2: Selection

- Selecting a subset
- Fictional literature in Swedish
  - bf:Text, bf:Monograph, bflc:encodingLevel
  - bf:genreForm (008/33)
- 1 million clusters → 85 000 clusters

National Library of Sweden

Step 3: Normalisation

- Copying properties within the cluster
  - bf:role
  - bf:contribution from bf:responsibilityStatement

BF2C217662AA	8ljb9wrq6pl6w2m0	s93qfmb44qxq6zk
	Title: jenny	Title: jenny
translationOf	Jenny • https://id.kb.se/language/nor	Jenny • https://id.kb.se/language/nor
genreForm	https://id.kb.se/marc/FictionNotFurtherSpecified	https://id.kb.se/marc/FictionNotFurtherSpecified
@type	Text	Text
language	https://id.kb.se/language/swe	https://id.kb.se/language/swe
_numPages	339	358
classification	kssb, 6: Hcedb.01	kssb, 6: Hcedb.01 kssb, 8: Hdb.01=c
contentType	https://id.kb.se/term/rda/Text	
contribution	aut: Sigrid, Undset, 1882-1949	aut: Sigrid, Undset, 1882-1949 trl: Sigrid, Elmblad, 1860-1926
reproductionOf		
instance title	[{@type=Title, mainTitle=Jenny}]	[{@type=Title, subtitle=roman, mainTitle=Jenny}]
instance type	Electronic	Instance
editionStatement		
responsibilityStatement	Sigrid Undset : roman / översättning av Sigrid Elmblad	av Sigrid Undset ; övers. av Sigrid Elmblad
encodingLevel	marc:AbbreviatedLevel	marc:AbbreviatedLevel
publication	1920 • Norstedt • Stockholm • https://id.kb.se/country/sw	1928 • Vårt hem • Stockholm • https://id.kb.se/country/sw
identifiedBy	kb-digitization-003076327	
extent	339 s.	358 s.
physicalDetailsNote		

National Library of Sweden

Step 4: Extraction

- Properties that prevent extraction
  - bf:contribution, bf:content, bf:intendedAudience, bf:translationOf
- Definition of Work
- Analyse the clusters

National Library of Sweden

Step 4: Extraction

94311F6D2201	q71m7bs24g2752g	6ph3606j0bkhfgj	3ld8w5kf3lklvz0	6phc076j4wqll7z
	Title: raddaren i noden	Title: raddaren i noden	Title: raddaren i noden	Title: raddaren i noden
@type	Text	Text	Text	Text
language	https://id.kb.se/language/swe	https://id.kb.se/language/swe	https://id.kb.se/language/swe	https://id.kb.se/language/swe
translationOf	The catcher in the rye • 4 • https://id.kb.se/language/eng	The catcher in the rye • 4 • https://id.kb.se/language/eng	The catcher in the rye • 4 • https://id.kb.se/language/eng	The catcher in the rye • 4 • https://id.kb.se/language/eng
genreForm	https://id.kb.se/marc/Novel https://id.kb.se/term/saogf/Romaner https://id.kb.se/term/saogf/Utvecklingsromaner	https://id.kb.se/marc/FictionNotFurtherSpecifi	https://id.kb.se/marc/Novel https://id.kb.se/term/saogf/Romaner https://id.kb.se/term/saogf/Utvecklingsromaner	https://id.kb.se/marc/Novel https://id.kb.se/term/saogf/Romaner https://id.kb.se/term/saogf/Utvecklingsromaner
subject	Caulfield • Holden • (fiktiv gestalt) https://id.kb.se/term/sao/F%C3%B6renta%20staterna • New York • Manhattan • https://id.kb.se/term/sao https://id.kb.se/term/sao/Flykt https://id.kb.se/term/sao/Ton%C3%A5rspojkar		Caulfield • Holden • (fiktiv gestalt) https://id.kb.se/term/sao/F%C3%B6renta%20staterna • New York • Manhattan • https://id.kb.se/term/sao https://id.kb.se/term/sao/Flykt https://id.kb.se/term/sao/Ton%C3%A5rspojkar	Caulfield • Holden • (fiktiv gestalt) https://id.kb.se/term/sao/F%C3%B6renta%20staterna • New York • Manhattan • https://id.kb.se/term/sao https://id.kb.se/term/sao/Flykt https://id.kb.se/term/sao/Ton%C3%A5rspojkar
_numPages	189	194	201	201
classification	kssb, 5: He=c	kssb, 8: Heq.01=c	kssb, 6: He.01=c kssb, 6: Hcee.01	kssb, 6: He.01=c
contentType				
contribution	aut: J. D., Salinger, 1919-2010 trl: Birgitta, Hammar, 1912-2011	aut: J. D., Salinger, 1919-2010 trl: Birgitta, Hammar, 1912-2011	aut: J. D., Salinger, 1919-2010 trl: Klas, Östergren, 1955-	aut: J. D., Salinger, 1919-2010 trl: Klas, Östergren, 1955-
reproductionOf				
instance title	[{@type=Title, mainTitle=Räddaren i nöden}]	[{@type=Title, mainTitle=Räddaren i nöden}]	[{@type=Title, mainTitle=Räddaren i nöden}]	[{@type=Title, subtitle=[roman], mainTitle=Räddaren i nöden}]
instance type	Instance	Instance	Instance	Print
editionStatement	[Ny utg.], 9. uppl.	2. uppl.	[Ny utg.]	
responsibilityStatement	J.D. Salinger ; övers. av Birgitta Hammar	J.D. Salinger ; övers. av Birgitta Hammar	J.D. Salinger ; översättning av Klas Östergren	J.D. Salinger ; översättning av Klas Östergren
encodingLevel	marc:FullLevel	marc:AbbreviatedLevel	marc:FullLevel	marc:FullLevel
publication	1982 • Bonnier • Stockholm • https://id.kb.se/country/sw	1964 • Aldus/Bonnier • Stockholm • https://id.kb.se/country/sw	1990 • Bonnier • Stockholm • https://id.kb.se/country/sw	1987 • Bonnier • Stockholm • https://id.kb.se/country/sw
identifiedBy	32:00		9100479039 • 30:70	9100473952 • INB. • 139:40
extent	189, [1] s.	194, [1] s.	201 s.	201 s.

National Library of Sweden

Step 4: Extraction

- Release December 2023
- 85 000 clusters → 48 000 Work entities (125 000 instances)
- Each Work linked to up to 60 instances
- Work properties merged/retained
- All instances linked to the stand-alone Work


```
"@id": "https://libris.kb.se/p719rbl11qn9b2m#it",
"@type": ”Print”,
"extent": [
   {
      "@type": "Extent",
      "label": [
         "239 s."
      ]
   }
],
"identifiedBy": [
   {
      "@type": "ISBN",
      "value": "9789174293814"
   }
],
 "instanceOf": {
"@id": "https://libris-qa.kb.se/6plmtjs54vwlgjt3#it"
},
 "issuanceType": "Monograph",
 "hasDimensions": {
    "@type": "Dimensions",
    "label": [
       "22 cm"
```


National Library of Sweden

Challenges

- LRM Expression
- Work title
  - bf:expressionOf.title → bf:title, bf:translationOf.title, bf:hasPart.title
- bf:Hub
- Illustrated works
  - Comics, children’s picture books
  - bf:role rdfs:domain
- MARC21 conversion
- Coexisting conceptual models

National Library of Sweden

Project team

1 Team Leader

3 Systems Developers

3 Systems Librarians

2 RDA Specialists

National Library of Sweden

Links

- Github: https://github.com/libris/librisxl/tree/develop/librisworks
- KBV Application Vocabulary: https://id.kb.se/vocab/
- Libris KBV Editor: https://libris.kb.se/katalogisering/

National Library of Sweden