Indexing German Audiovisual Heritage - SKOSification Made Easy

Johannes Hercher & Dr. Harald Sack

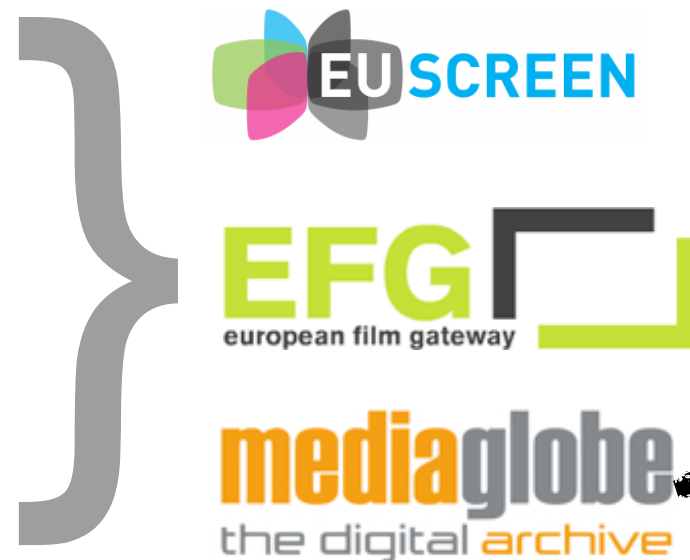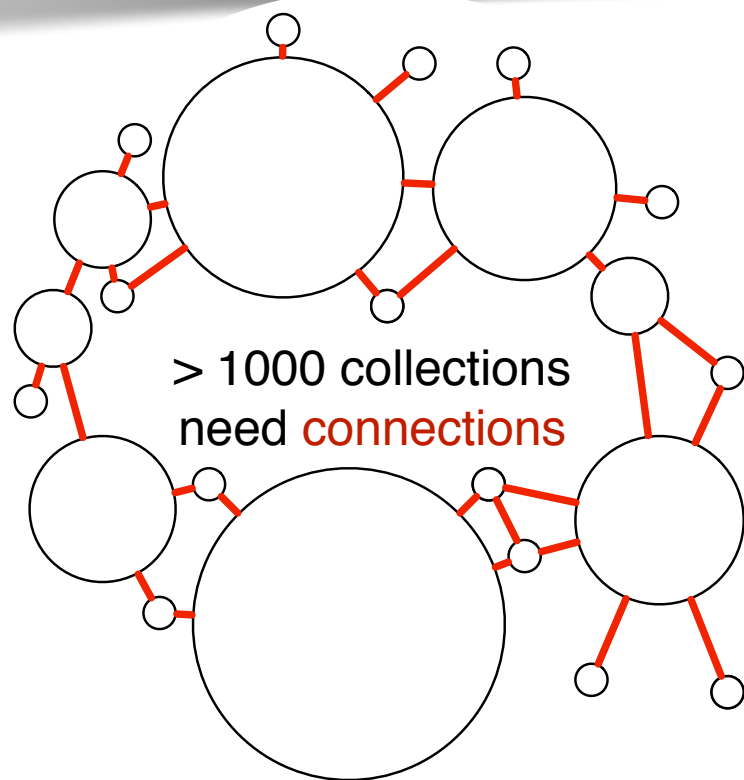Hasso-Plattner-Institute for IT Systems Engineering

Potsdam, Germany

What's the Problem?

Cultural heritage taped on film (Germany)

> 1.200 film collections*

> 200 institutions with more than 100 h of film**

~ 83% of film archives plan to digitise their collections

Classification & Thesauri needed:

☐ **integrate** various collections

☐ **harness** semantic relations

☐ **align** content

> 1000 collections need connections

EU SCREEN

EFG
european film gateway

mediaglobe
the digital archive

Semantic glue

Classifications and **Thesauri**

but...

!

vocabularies missing

*) Bockhorn, V; Morsbach; Ronneburg, S; Schmidt, W. (1996). Topographie audiovisueller Quellenüberlieferung - Film- und Videobestände in Archiven und archivischen Einrichtungen in der Bundesrepublik Deutschland. Koblen
**) Mediaglobe, Deliverable-1.1 Report 2010.

5

mediaglobe - the Digital Archive

makes cultural heritage accessible & tangible

- **augmenting workflows** in media-archives
  - □ efficient multimedia-analysis
  - □ (semi-)automated semantic indexing
  - □ retrieval of AV-content

- **covers**
  - □ >1.000h of AV-data (defa-spektrum Archive)
  - □ newsreels, interviews, factual recordings, …

<····· broader ····> narrower ·····>

Federal Ministry of Economics and Technology

THESEUS
New Technologies for the Internet of Services

mediaglobe
the digital archive

defa spektrum

The Need for Vocabularies

- **AV-media** needs **descriptions**

  - ☐ content topics

  - ☐ types of broadcast

  - ☐ genres

  - ☐ …

flower flowers food france friends fun garden geot
germany girl graffiti green halloween hawaii hiking holiday
honeymoon hongkong house india ireland island italy japan july
lake landscape light live london losangeles macro me mexico

tagclouds **VS** thesauri

- **tags** are **not sufficient**…

  - ☐ variant, ambiguous, flat

  - ☐ => inconsistent, uncontrolled

- **controlled vocabulary** *is* *suited*
  due to **term-relations**

TT Films
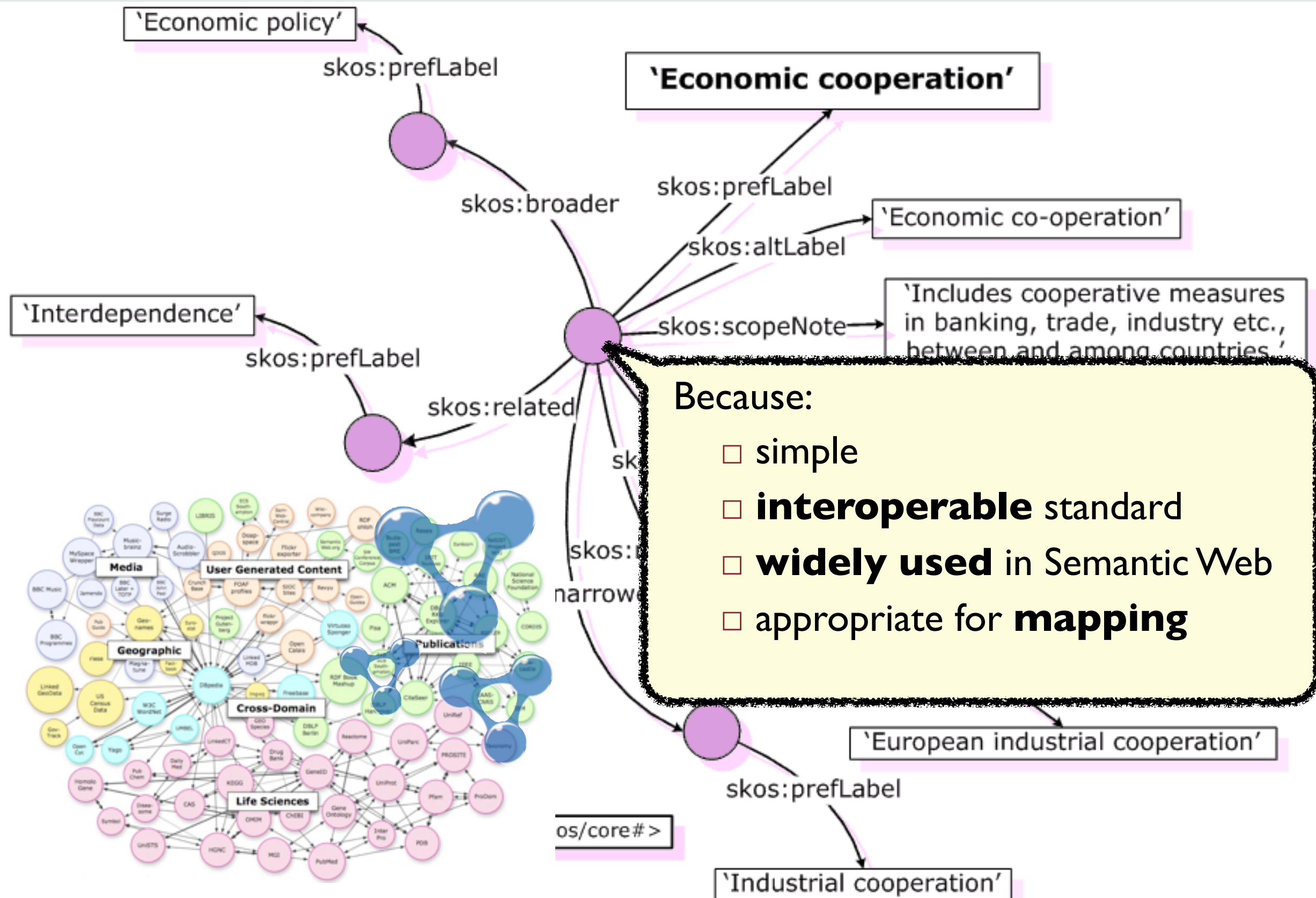   UF Motion picture
    NT Nonfiction films
   (...)
        NT Documentary films
            SN A heading used for ...
            UF Documentaries, Motion picture
            UF Factual films
            UF Motion pictures documentaries
            BT Nonfiction films
            RT Actualities (Motion pictures)

'Economic policy'

skos:prefLabel

'Economic cooperation'

skos:broader

skos:prefLabel

skos:altLabel → 'Economic co-operation'

'Interdependence'

skos:scopeNote → 'Includes cooperative measures in banking, trade, industry etc., between and among countries.'

skos:prefLabel

skos:related

Because:

☐ simple

☐ **interoperable** standard

☐ **widely used** in Semantic Web

☐ appropriate for **mapping**

narrow

'European industrial cooperation'

os/core#>

skos:prefLabel

'Industrial cooperation'

Backing Problem Statement

3.1 How to **Find Vocabularies**

- **Catalogues**

  - ☐ http://www.w3.org/2001/sw/wiki/SKOS/Datasets

  - ☐ http://www.vocabularyserver.com

  - ☐ http://www.taxonomywarehouse.com/

  - ☐ …

- **Specialised search engines**

- **Major associations**

3.2 **Vocabularies Related to AV-Media**

- **Results**

  - ☐ AV-related Vocabularies are mainly in english language.

  - ☐ German broadcasters do NOT publish vocabularies !

  - ☐ Many small and mid-sized archives have own *classifications*.

  - ☐ **Low availability** of vocabularies (in german language).

	Name	SKOS
IPTC	Thesaurus International Press Telecommunication Council	X
TGM	Thesaurus for Graphic Materials (Library of Congress)	X
FIAF GSH	FIAF General Subject Headings	X
MIM	Moving Image Materials	X
LCGFT	Moving Image Genre-Form Headings (Library of Congress)	(X)
T-POL 🇩🇪	Thesaurus Politics of (former) GDR-Filmarchive (~12.000 Terms)	-
REM 🇩🇪	~6 co-existing thesauri of German broadcasting compound.	-
Small Archives 🇩🇪	>200 Collections with custom classification. Only a few use thesauri-terms of the public broadcasting compound.	-

- **appropriate vocabularies** exist…

- But we can not use them, because

  - they are not encoded in RDF

  - they are not published

  - (some) are not up to date

some answers from telephone interviews

"… our vocabulary needs review and revision first…"

"… ownership and copyrights not clear at the moment…"

"… we are **interested but** have **no resources**…"

Our Proposal

»**SKOSify existing vocabularies** (even more) **easily**

■ **Main requirements**

☐ simple ingest (example follows)

☐ download/publish an SKOS/RDF encoded vocabulary

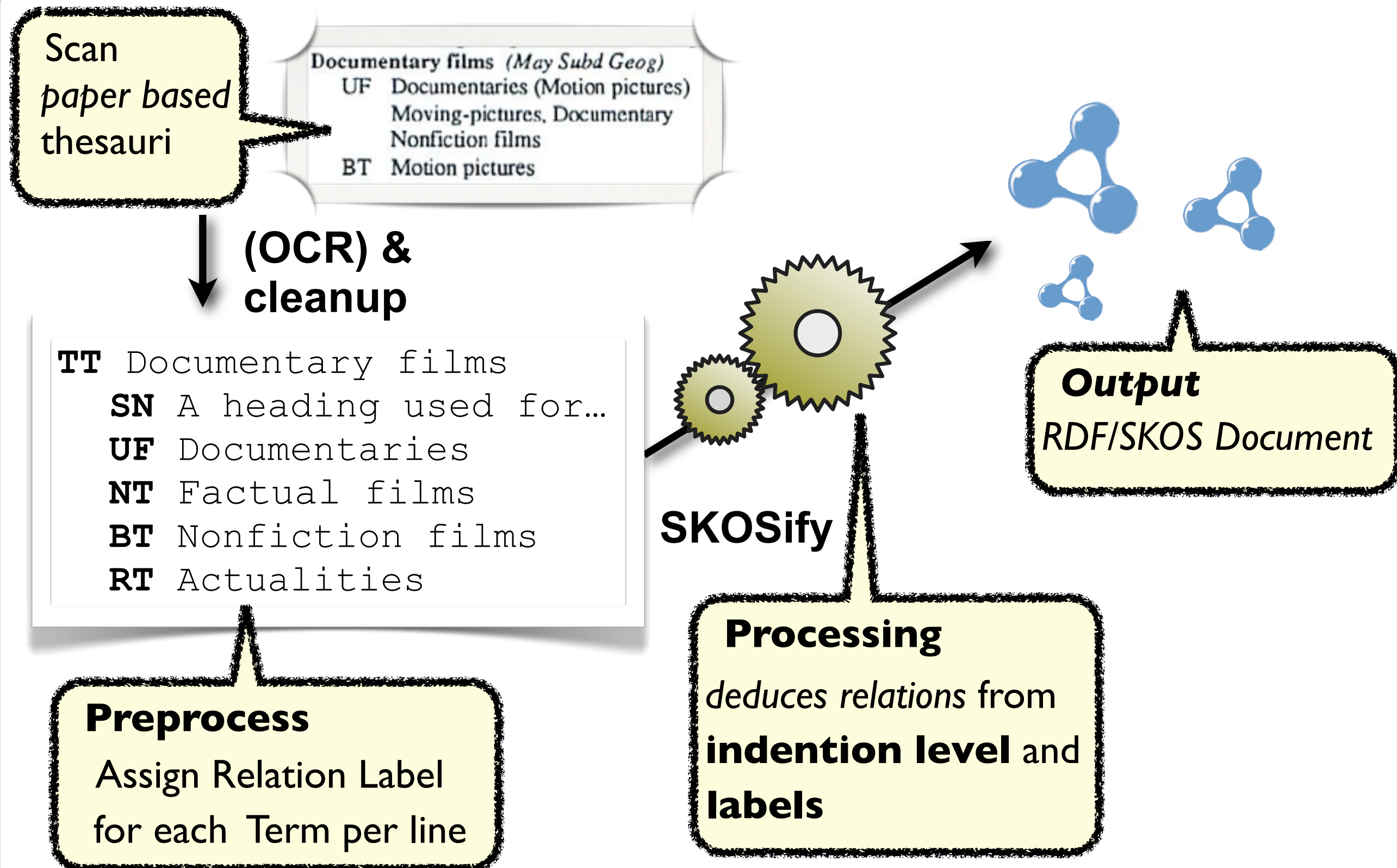☐ low learning curve (simple)

☐ no installation required

■ **Need more features?**

☐ indexing assistance & term search

☐ mapping & collaborative development

☐ management & visualisation

}

xTree
(digicult)

Scan *paper based* thesauri

> **Documentary films** *(May Subd Geog)*
> UF   Documentaries (Motion pictures)
>   Moving-pictures, Documentary
>   Nonfiction films
> BT   Motion pictures

(OCR) & cleanup

```
TT Documentary films
   SN A heading used for...
   UF Documentaries
   NT Factual films
   BT Nonfiction films
   RT Actualities
```

SKOSify

Output
RDF/SKOS Document

Preprocess
Assign Relation Label for each Term per line

Processing
deduces relations from **indention level** and **labels**

quickSKOS

http://skosifier.org

Name your vocabulary.

thesaurus on films (?)

< Input > (?)

TT Film genre
  NT Animated films
  NT Documentary films
    SN This is a heading used for ...
    UF Documentaries, Motion picture
    UF Factual films
    UF Motion picture documentaries
    BT Nonfiction films
    RT Actualities (Motion pictures)
  NT Filmed plays
  ...
TT Film storages

HOW TO

1. Name your vocabulary.
2. Describe it using labels.
3. Determine your namespace.
4. Provide general information.
5. Download or publish it at SKOSifier.

free of charge hosting of your vocabulary
at skosifier.org ?

Do you need hosting? (?)

○ yes, please   ○ no, thanks

edit your namespace

if hosting "no" selected,
put in your own namespace

http://www.example.com/yourNamespace/ (?)

general information

such as scope, version and authors

... (?)

Download as SKOS   Publish at skosifier

Questions

Comments

Discussion

Thank You!

Johannes Hercher
phone: +49 (0)331-5509-547
email: Johannes.Hercher@hpi.uni-potsdam.de

Dr. Harald Sack
phone: +49 (0)331-5509-527
email: Harald.Sack@hpi.uni-potsdam.de

http://www.hpi.uni-potsdam.de/meinel/forschung/future_internet/semantic_technologies.html

http://www.projekt-mediaglobe.de/