2013 DCMI-AsiaPac Workshop on RDA, DC and Linked Data

# NLK Linked Data (LD) Publishing Experience


Sam Oh
Professor, Sungkyunkwan University, LIS
Affiliate Professor, University of Washington
ISO/IEC JTC1/SC34 Chair
ISO TC46/SC9 Chair
DCMI Oversight Committee Member

Jinho Park
Senior Researcher, National Library of Korea
ISO/TC46 SC9 Korea Secretariat
Member of the International Relations Committee, Korean Library Association
CJKDLI Korea Working Group Leader

# 3. NLK-LD Strategy

- NLK's decision to open its data as LD
  - NLK data is public data that should be available to everyone
  - The NLK must be open and easily accessible to others
- Global standard formats for opening NLK data
  - Respecting LD principles when opening is recommended
- Adding to the Global Information Ecosystem
  - The Web is the most general and accessible platform and ecosystem
  - Contributing to the Web is of the utmost importance

# 3. NLK-LD Strategy

- Focused more on contributing to the global database (Web) rather than what the NLK has to gain

- Emphasis was placed on increasing users' data literacy

# 3. NLK Linked Data Strategy

- The final goal is to establish a library data platform

- The platform should
  - Provide easy ways to use data regardless of their formats (MARC, OAI, Open API, RDF, Odata, JSON, XML)
  - Allow one to search ‘vocabulary’ or ‘data models’ and use them seamlessly (FOAF, FRBR, SKOS, DC, SIOC, MODS, PREMIS, BIBFRAME, EDM)

# NLK-LD Project

- Publishing NLK Data as LD started in 2013
  - 2011 : Preliminary research performed to establish a general understanding of and strategies for NLK Linked Data
  - 2012 : Partial data (bibliography, subject heading, name authority) converted to RDF and analysis was carried out on the problems of KORMARC as raw data
  - 2013 : Project set in motion based on bibliographic, subject heading, and name authority data unrestricted by issues of copyright and privacy
    - KORMARC2RDF, DB2RDF
    - Building a LD Platform (Data inquiry, download, SPARQL etc.)
    - Building a LD Management System (Interlinking management, etc.)
    - Development of applied services for general users

# Thesis on NLK-LD

- Jinho Park, “A Study on the Expansion of Information Integration through Linked Data Conversion and Library Data”
  - To identify integrated linkage, expansion through linked data conversion was presented in a theoretical manner and external linked data occurring in a practical manner was obtained by converting large-scale bibliographic, name authority, and subject heading data converted by the NLK into Linked Data types

# Research Questions

- The following four research questions were raised.
  - First, what factors make it difficult to convert bibliographic, subject heading, and name authority data into Linked Data types?
  - Second, what factors make it difficult to create detailed connections when bibliographic information, subject headings, and name authority data are converted in accordance with Linked Data principles?
  - Third, what information integration of raw data expanded when bibliographic data was converted into Linked Data for interlinking?
  - Fourth, what benefits and technical factors are offered through Graphic User Interface data management and interlinking management of Linked Data?

# Current Test Data

- Current Raw Data

| Classification | Extraction Method | Cases/Size Extracted |
|---|---|---|
| Bibliographic Book Data | Editable KORMARC format | 3,781,809 cases |
| Name Authority Data | Editable KORMARC format | 213,548 cases |
| Subject Heading Data | MSSQL mdb file format | 370,464 KB |

- The TERM and THESAURUS tables were key to the subject heading database
- A total of 560,561 subject headings in the TERM table
- A total of 1,600,637 word relations in the THESAURUS table

# Current Source Data Mapping

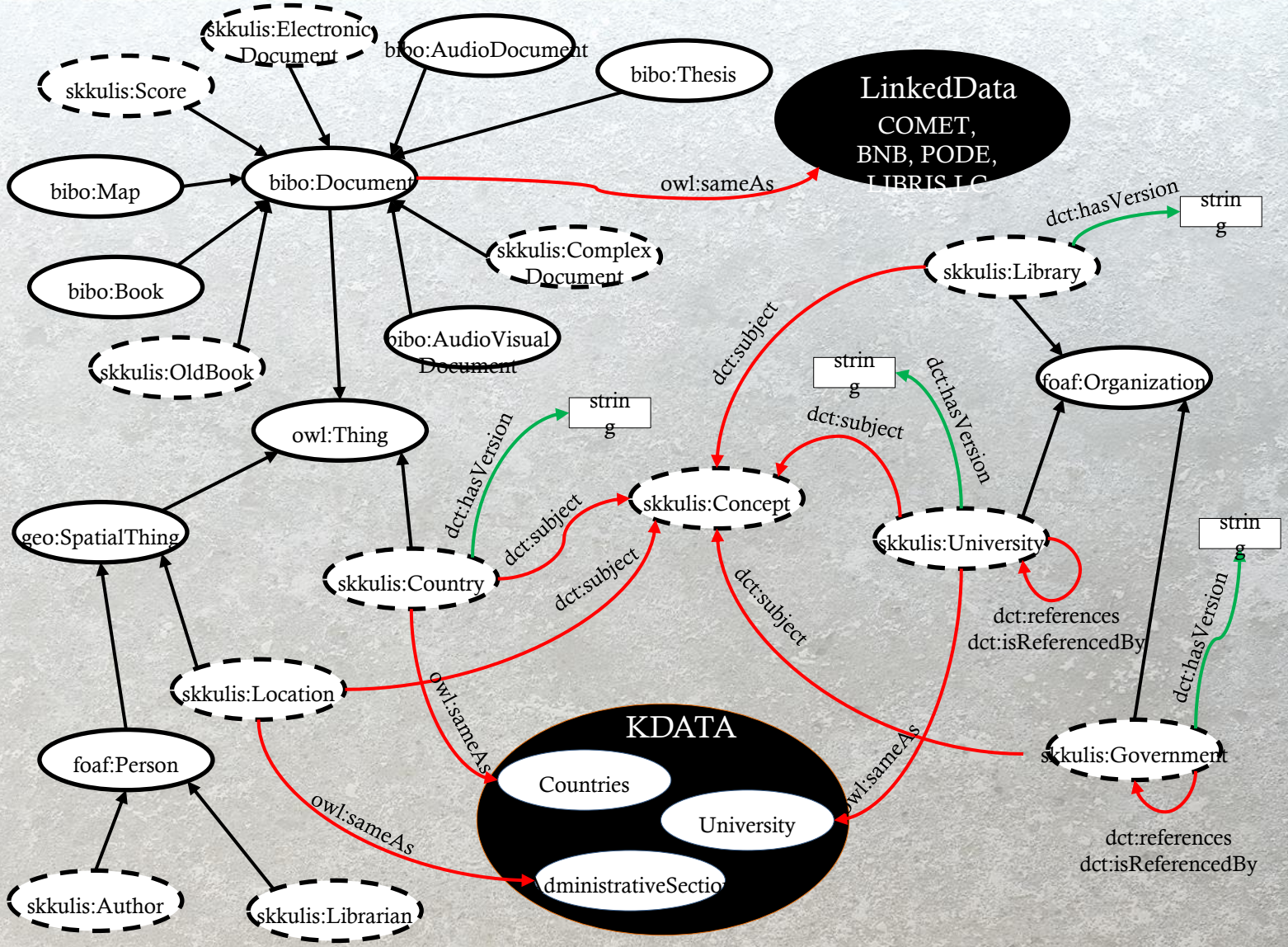
- Vocabulary Mapping

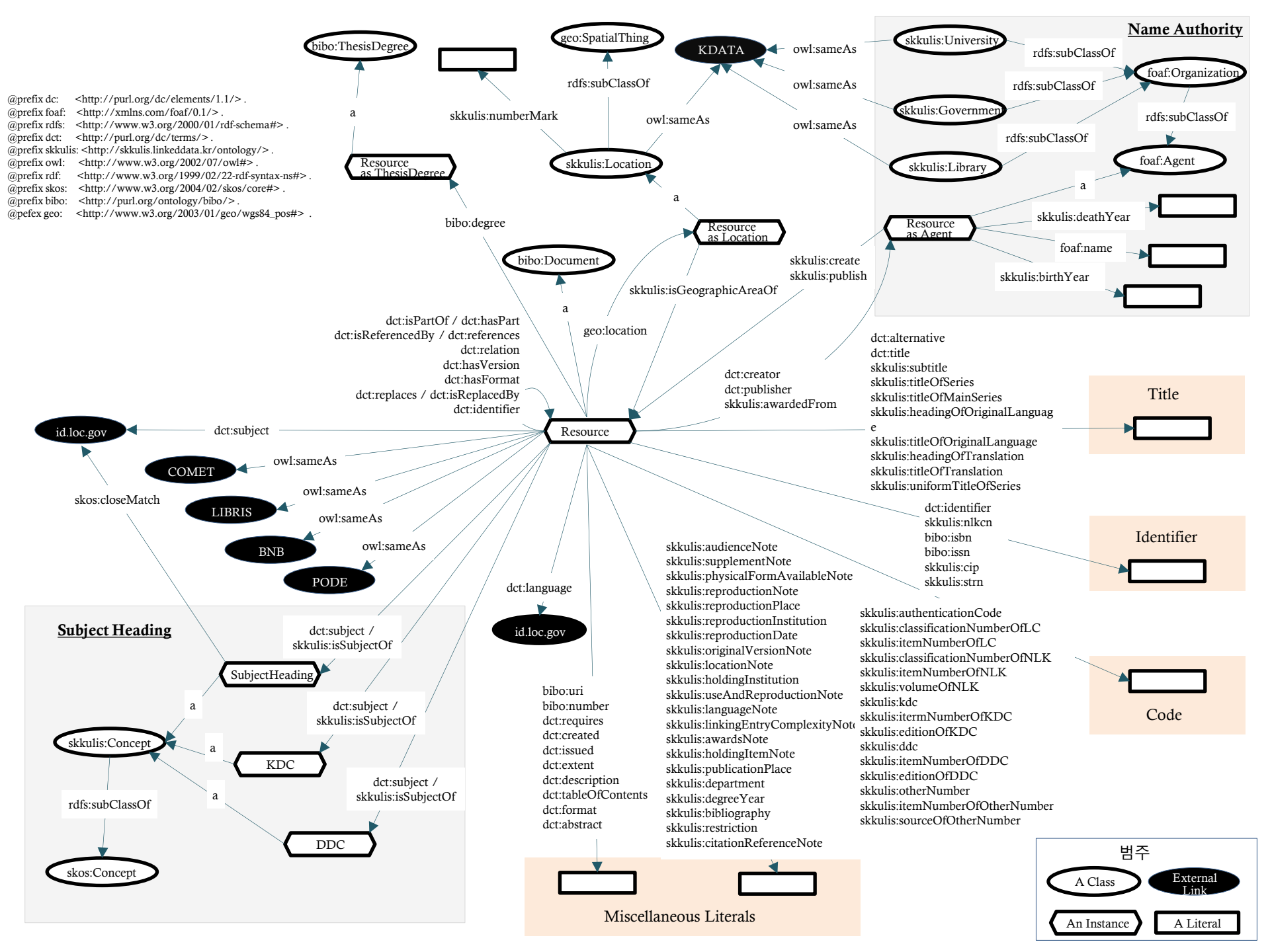
| Classification | Target Vocabulary |
|---|---|
| Bibliographic Book Data | BIBO, DC |
| Name Authority Data | FOAF |
| Subject Heading Data | SKOS |

# KORMARC2RDF Conversion

| KORMARC | | RDF | | |
|---|---|---|---|---|
| Field | Subfield | Subject Type | Predicate | Object Type |
| 245 | a | bibo:Document | dct:title | rdfs:Literal |
| | b | bibo:Document | skkulis:subtitle | rdfs:Literal |
| 260 | a | bibo:Document | skkulis:publicationPlace | rdfs:Literal |
| | b | bibo:Document | dct:publisher<br>skkulis:publish(inverse) | foaf:Organization |
| | c | bibo:Document | dct:created | rdfs:Literal |
| | g | bibo:Document | dct:issued | rdfs:Literal |
| 300 | a,b,c,e | bibo:Document | dct:extent | rdfs:Literal |
| 490 | s | bibo:Document | dct:isPartOf<br>dct:hasPart(inverse) | bibo:Document |

# Class Diagram



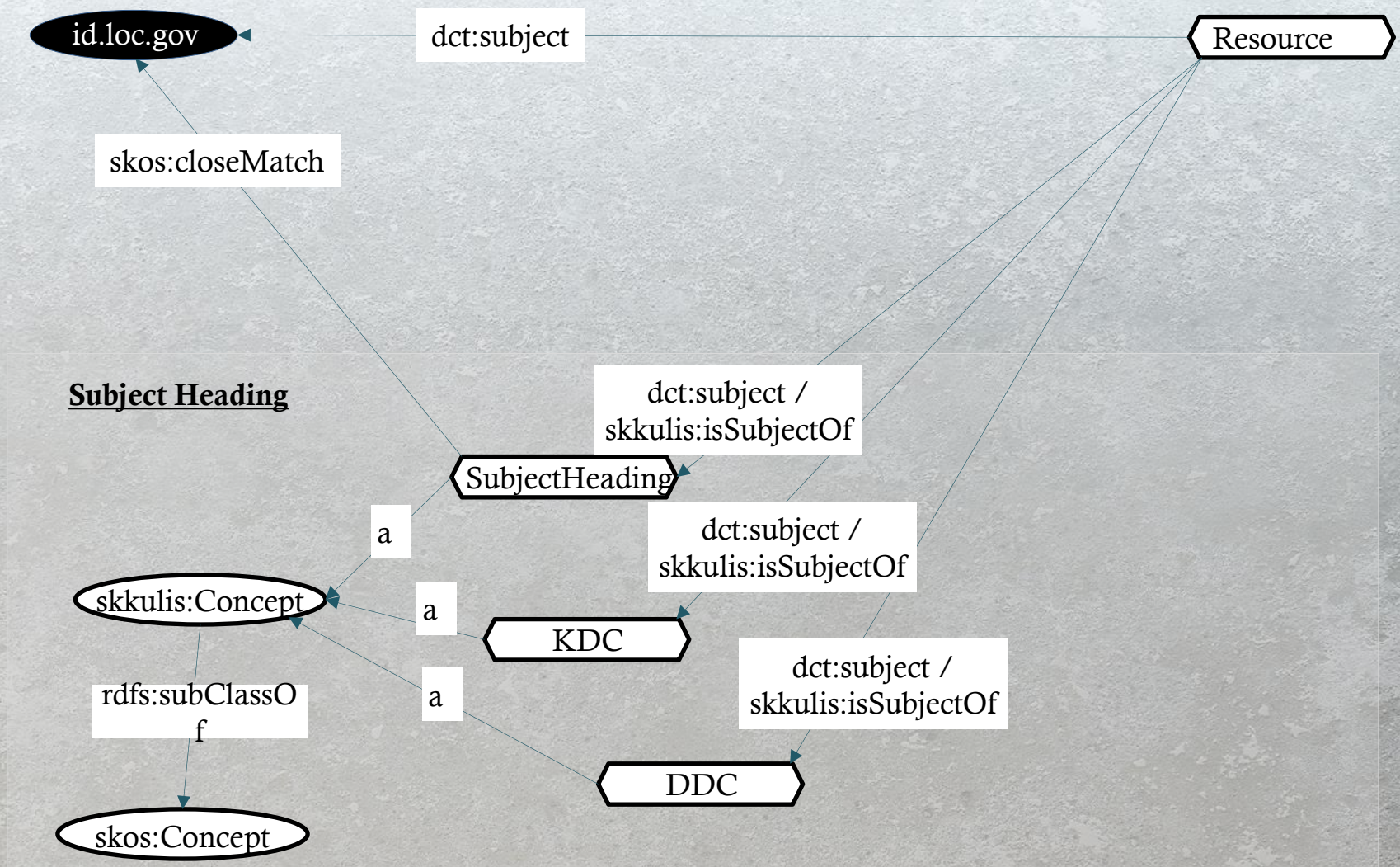

@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix skkulis: <http://skkulis.linkeddata.kr/ontology/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix bibo: <http://purl.org/ontology/bibo/> .
@pefex geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
bibo:ThesisDegree
geo:SpatialThing
KDATA
owl:sameAs
owl:sameAs
owl:sameAs
a
skkulis:numberMark
rdfs:subClassOf
owl:sameAs
Resource as ThesisDegree
skkulis:Location
a
bibo:degree
Resource as Location
bibo:Document
a
skkulis:isGeographicAreaOf
geo:location
skkulis:create
skkulis:publish
Name Authority
skkulis:University
rdfs:subClassOf
foaf:Organization
rdfs:subClassOf
skkulis:Government
rdfs:subClassOf
rdfs:subClassOf
skkulis:Library
foaf:Agent
a
Resource as Agent
skkulis:deathYear
foaf:name
skkulis:birthYear
dct:isPartOf / dct:hasPart
dct:isReferencedBy / dct:references
dct:relation
dct:hasVersion
dct:hasFormat
dct:replaces / dct:isReplacedBy
dct:identifier
dct:creator
dct:publisher
skkulis:awardedFrom
Resource
id.loc.gov
dct:subject
COMET
owl:sameAs
LIBRIS
owl:sameAs
BNB
owl:sameAs
PODE
owl:sameAs
skos:closeMatch
dct:alternative
dct:title
skkulis:subtitle
skkulis:titleOfSeries
skkulis:titleOfMainSeries
skkulis:headingOfOriginalLanguage
skkulis:titleOfOriginalLanguage
skkulis:headingOfTranslation
skkulis:titleOfTranslation
skkulis:uniformTitleOfSeries
Title
dct:identifier
skkulis:nlkcn
bibo:isbn
bibo:issn
skkulis:cip
skkulis:strn
Identifier
dct:language
id.loc.gov
skkulis:audienceNote
skkulis:supplementNote
skkulis:physicalFormAvailableNote
skkulis:reproductionNote
skkulis:reproductionPlace
skkulis:reproductionInstitution
skkulis:reproductionDate
skkulis:originalVersionNote
skkulis:locationNote
skkulis:holdingInstitution
skkulis:useAndReproductionNote
skkulis:languageNote
skkulis:linkingEntryComplexityNote
skkulis:awardsNote
skkulis:holdingItemNote
skkulis:publicationPlace
skkulis:department
skkulis:degreeYear
skkulis:bibliography
skkulis:restriction
skkulis:citationReferenceNote
skkulis:authenticationCode
skkulis:classificationNumberOfLC
skkulis:itemNumberOfLC
skkulis:classificationNumberOfNLK
skkulis:itemNumberOfNLK
skkulis:volumeOfNLK
skkulis:kdc
skkulis:itermNumberOfKDC
skkulis:editionOfKDC
skkulis:ddc
skkulis:itemNumberOfDDC
skkulis:editionOfDDC
skkulis:otherNumber
skkulis:itemNumberOfOtherNumber
skkulis:sourceOfOtherNumber
Code
Subject Heading
dct:subject / skkulis:isSubjectOf
SubjectHeading
a
skkulis:Concept
a
KDC
dct:subject / skkulis:isSubjectOf
a
DDC
dct:subject / skkulis:isSubjectOf
rdfs:subClassOf
skos:Concept
bibo:uri
bibo:number
dct:requires
dct:created
dct:issued
dct:extent
dct:description
dct:tableOfContents
dct:format
dct:abstract
Miscellaneous Literals
범주
A Class
External Link
An Instance
A Literal

# Results of Source Data Conversion

| Classification | Number of Target Conversions | Number of RDF Conversions |
|---|---|---|
| Bibliographic Book Data | 3,781,809 | 87,561,248 |
| Name Authority Data | 560,561 | 8,018,182 |
| Subject Heading Data | 213,548 | 1,631,255 |

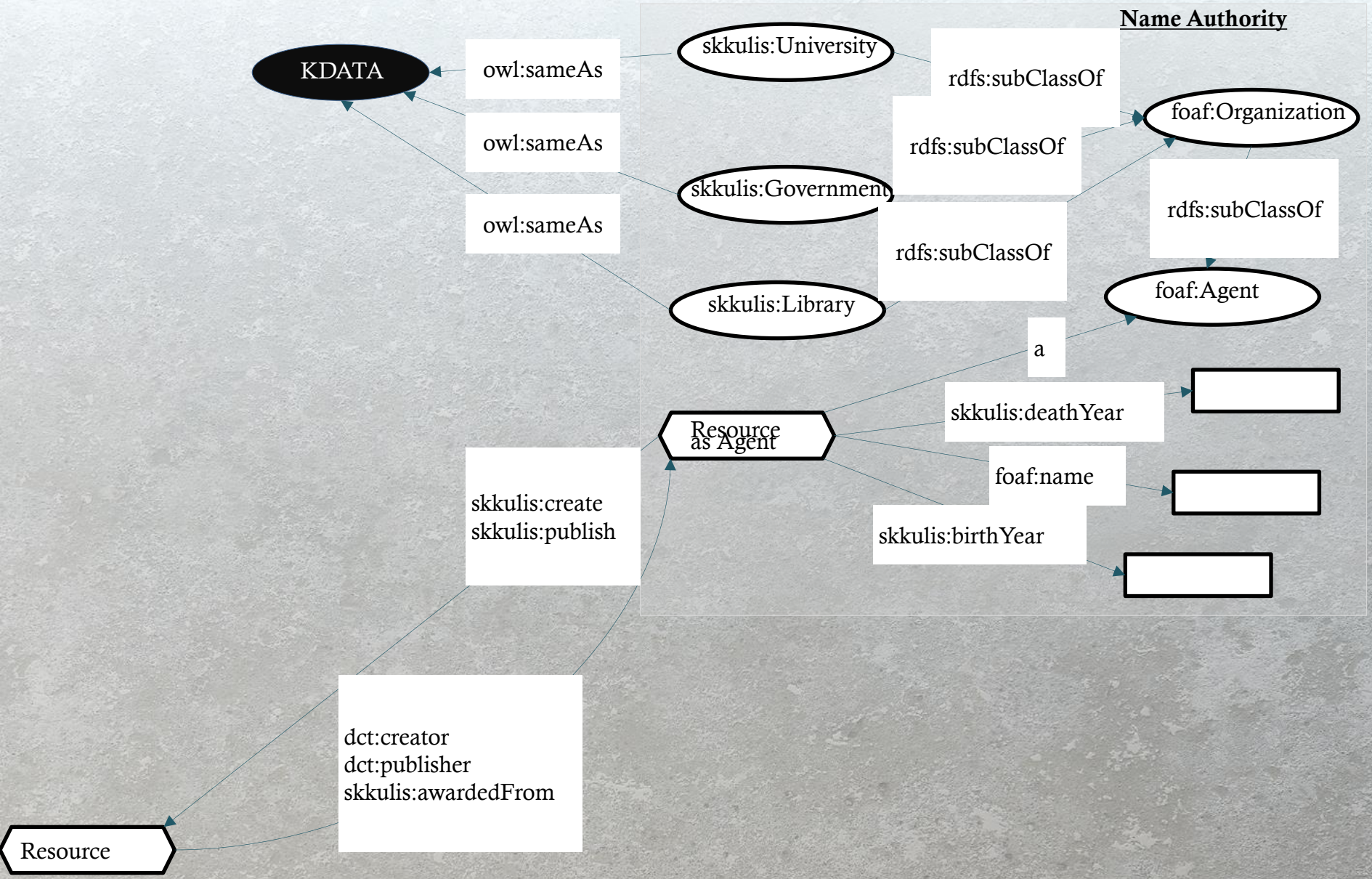
# Interlinking within NLK Data

- Bibliographic Data and Subject Headings
  - Utilized subject headings recorded in the 6XX added entries within the bibliographic data
  - Selection of subject headings matching the KDC and DDC


id.loc.gov
dct:subject
Resource
skos:closeMatch
Subject Heading
dct:subject / skkulis:isSubjectOf
SubjectHeading
a
dct:subject / skkulis:isSubjectOf
skkulis:Concept
a
KDC
dct:subject / skkulis:isSubjectOf
rdfs:subClassOf
a
DDC
skos:Concept

# Interlinking within NLK Data

- Bibliographic Data and Name Authority
  - Linked control number of name authority data in the bibliographic data with the control number of the name authority data
  - Target fixed as control numbers for name authorities within the 100, 700, 710 fields of the bibliographic data


Name Authority
KDATA
owl:sameAs
owl:sameAs
owl:sameAs
skkulis:University
skkulis:Government
skkulis:Library
rdfs:subClassOf
rdfs:subClassOf
rdfs:subClassOf
foaf:Organization
rdfs:subClassOf
foaf:Agent
a
Resource
as Agent
skkulis:deathYear
foaf:name
skkulis:birthYear
skkulis:create
skkulis:publish
dct:creator
dct:publisher
skkulis:awardedFrom
Resource

# Interlinking to External Data

- Attempt to link bibliographic data with COMET, BnB, LIBRIS, and PODE
- Attempt to link subject headings with subject headings within the U. S. Library of Congress
- Attempt to link name authorities with VIAF (failed)

# Interlinking to External Data

- Attempt to link bibliographic data with COMET, BnB, LIBRIS, and PODE
  - COMET, BNB, PODE provide SPARQL Endpoint with which the study was able to collect data to link to
    - ❖Searched for bibliographic data including ISBN(field 020) and ISSN(field 022) and formulated its control number and ISBN, ISSN numbers into text A
    - ❖Used SPARQL to search for triples including ISBN and ISSN within their endpoints and formulated subjects and objects into text B
    - ❖Matched text A with text B to test for identical ISBN and ISSN before formulating the control numbers of identical results into text M
    - ❖Converted and connected text M into triple (owl:sameAs)

# Interlinking to External Data

- Attempt to link bibliographic data with COMET, BnB, LIBRIS, and PODE
  - Data collected through the Xsearch API provided by LIBRIS
    - ❖Searched for bibliographic data including ISBN(field 020) and ISSN(field 022) and formulated its control number and ISBN, ISSN numbers into text A
    - ❖Checked whether searching the ISBN or ISSN numbers in text A through the factors of the LIBRIS search API returned any results before saving those results as text M
    - ❖Converted and connected text M into triple (owl:sameAs)

# Interlinking to External Data

- Attempt to link subject headings with subject headings within the U. S. Library of Congress
  - Downloaded subject heading data from the U. S. Library of Congress in the form of .nt and connected it with identical subject headings within NLK data
    - Searched for fields with subject headings within the bibliographic data and formulated their control numbers and subject headings into text S
    - Downloaded the Library of Congress subject file and formulated its subject heading control numbers and subject headings into text L
    - Compared and checked text S and text L for identical subject headings
    - Converted and connected identical subject headings from text S into triple (dct:subject)

# Results of Bibliographic Data Linkage

| | NLK Data | COMET Linkage Results | | BNB Linkage Results | | LIBRIS Linkage Results | | PODE Linkage Results | |
|---|---|---|---|---|---|---|---|---|---|
| ISBN | 1,632,214 | **25,088** | **1.537%** | X | | | | **111** | **0.007%** |
| ISBN10 | X | X | | **76,409** | **4.681%** | **74,663** | **18.959 %** | X | |
| ISBN13 | X | X | | **3,667** | **4.681%** | | | X | |
| ISSN | 6,224 | X | | **1,180** | **18.959%** | **3,579** | **57.503 %** | 0 | |

- Results of linkage with external data seem slight when applying Korean-centered bibliographic book data from the NLK as a standard

# Results of Linkage by Target Data Type

## Data as Categorized in KORMARC Leader

| Code | Classification | Number of Types Linked | Percentage within Linked Data | Number of Source Data Types | Percentage of Linkage within Source Data |
|---|---|---|---|---|---|
| WMO | Western Books | 235,756 | 68.5321% | 322,530 | 73.0958% |
| WMF | Western Micro-Data | 78,303 | 22.7620% | 256,769 | 30.4955% |
| WJU | Western Children's Books | 17,947 | 5.2170% | 22,987 | 78.0746% |
| WDM | Western Dissertations | 3,304 | 0.9604% | 21,488 | 15.3760% |

# Results of Subject Heading Linkage

- Current Linkage between NLK Subject Headings and U.S. Library of Congress Subject Headings

| Classification | Number of Linkages | Number of NLK Subject Headings Linked | Number of U. S. Library of Congress Subject Headings Linked |
|---|---|---|---|
| Linkage of Bibliographic Data and Subject Headings | 794,697 | 318,297 | 43,965 |
| Linkage of Subject Headings | 10,245 | 9,695 | 10,245 |

# Difficult Factors in Interlinking

- Factors Complicating Internal Interlinking
  - The percentage of bibliographic data with a subject heading value (775,723 cases) amounts to 20.512% of the total bibliographic data (3,781,809 cases)
  - The percentage of bibliographic data with a subject heading value within the authority subject heading database (615,080 cases) amounts to 16.264% of the total bibliographic data (3,781,809 cases)
  - The number of subject heading terminology which surfaces in the bibliographic data amounts to 128,776 cases, 50,118 of which exist within the authority subject heading database and 78,658 of which were entered at random as a literal string
  - The percentage of subject headings actually put to system use (50,118 cases) amounts to 8.941% of the total number of subject headings within the NLK (560,561 cases)

# Difficult Factors in Interlinking

- Factors Complicating Internal Linking
  - The percentage of bibliographic data with a name authority value (1,312,435 cases) amounts to 34.704% of the total bibliographic data (3,781,809 cases)
  - The percentage of bibliographic data with a name authority control number (136,913 cases) amounts to 3.620% of the total bibliographic data (3,781,809 cases)
  - As in the case of the authority subject heading database, the name authority database also controls only 3.620% of terminology used
  - All other name authority data values were input as a literal string and therefore cannot be linked to bibliographic data

# Difficult Factors in Interlinking

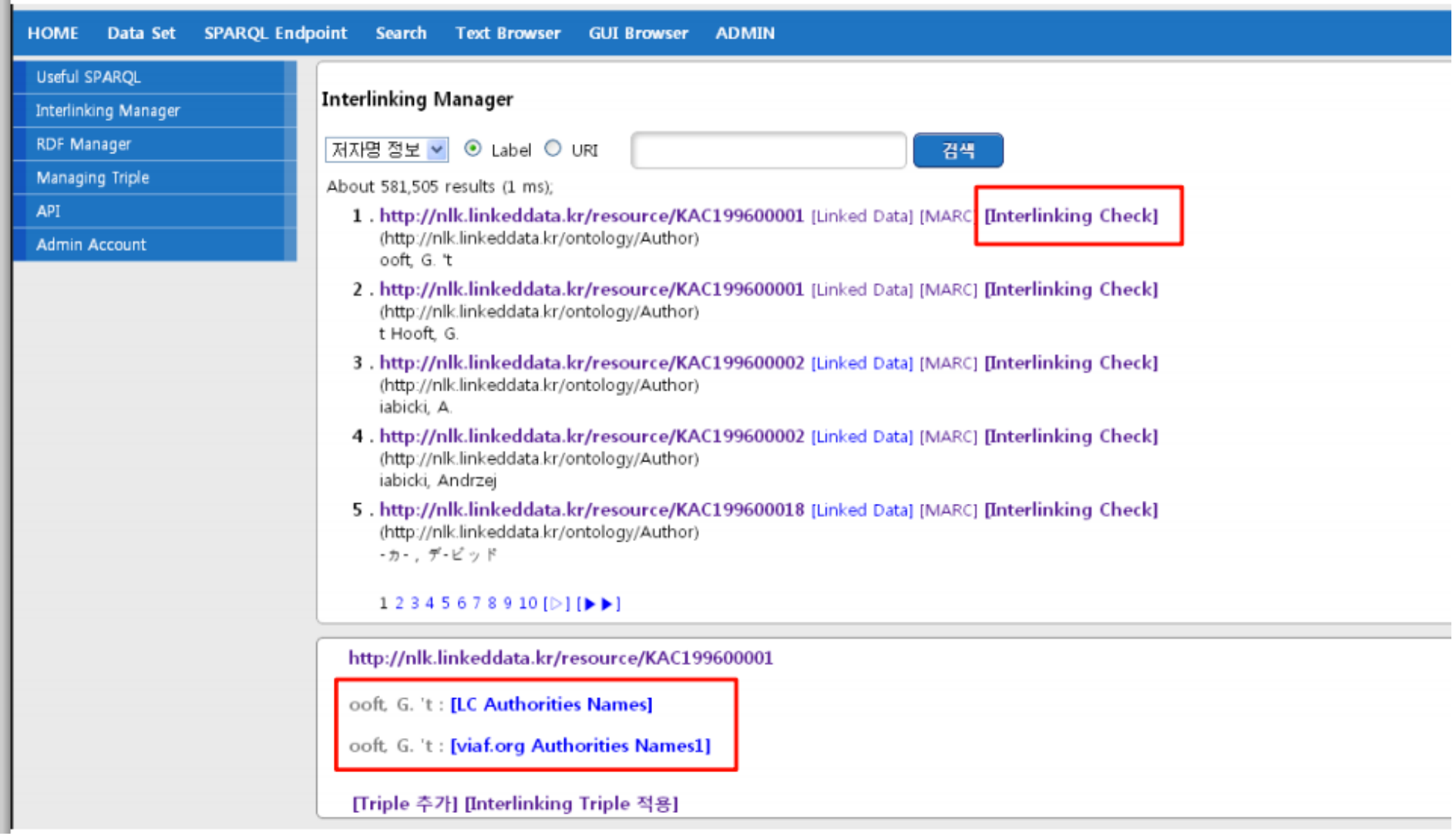
- Failed Attempt to Link Name Authority Data to VIAF
  - Unlike bibliographic data, name authority data is not distinctively differentiated by **international identifiers**
  - As in the case of subject headings, Korean terminology does not have corresponding English or French terminology, making it impossible to link matching entities
  - **Diverse methods of notation (i.e., Romanization)** for name authority entities further complicate linkage

# RDF Data Management Interface

- Building a GUI Environment for RDF Data and Interlinking Management

| Function | Description |
|---|---|
| SPARQL Endpoint | Provides data contact points via SPARQL |
| Integrated Search Function | Provides search capabilities for labels and URI concerning the entire body of data in addition to individual bibliographic, name authority, and subject heading data |
| Bibliographic Data Search Function | |
| Name Authority Search Function | |
| Subject Heading Search Function | Provides class type, label, and MARC data on resources turned up during search |
| **Interlinking Management** | **Provides interlinking checks for relevant data and contact points for external LOD** |
| RDF Management | Provides management capabilities for given data |

# Interlinking Management Example



# A GUI-Based LD Management System

- Conducted Interviews with Chief of Metadata and Link Data Developer at the NLK
  - Benefits of Visual LD Management Screen for Data Managers
    - Ability to search LD without specialized knowledge
    - Advanced LD management, including exploration of the relationship between resources
    - New deduction-based linkage management
    - Cutback on search time due to visual management of related information

# A GUI-Based LD Management Systems

- Conducted Interviews with Chief of Metadata and Link Data Developer at the NLK
  - Technical requirements for the Visual LD Management System
    - Query-less SPARQL interface
    - Internal and external vocabulary management module
    - Monitoring interface
    - External data set surveillance
    - Visual editing of ontologies and RDF
    - Interlinking monitoring
    - Formulation and management of diverse SPARQL examples