Linked Data Applications for Library Collections

•••

Elk Oswood

Linked Data and the Semantic Web

Semantic Web (SW) Project

- Tim Berners-Lee (2006 W3C)
- Extend Web to make data machine-readable

Linked Data

- Set of best practices for publishing data on the Web.

Talk Outline

1. Overview of Relevant Web Technologies
2. Overview of Linked Data
3. Applications of Linked Data for Library Use
4. Conclusion and Discussion

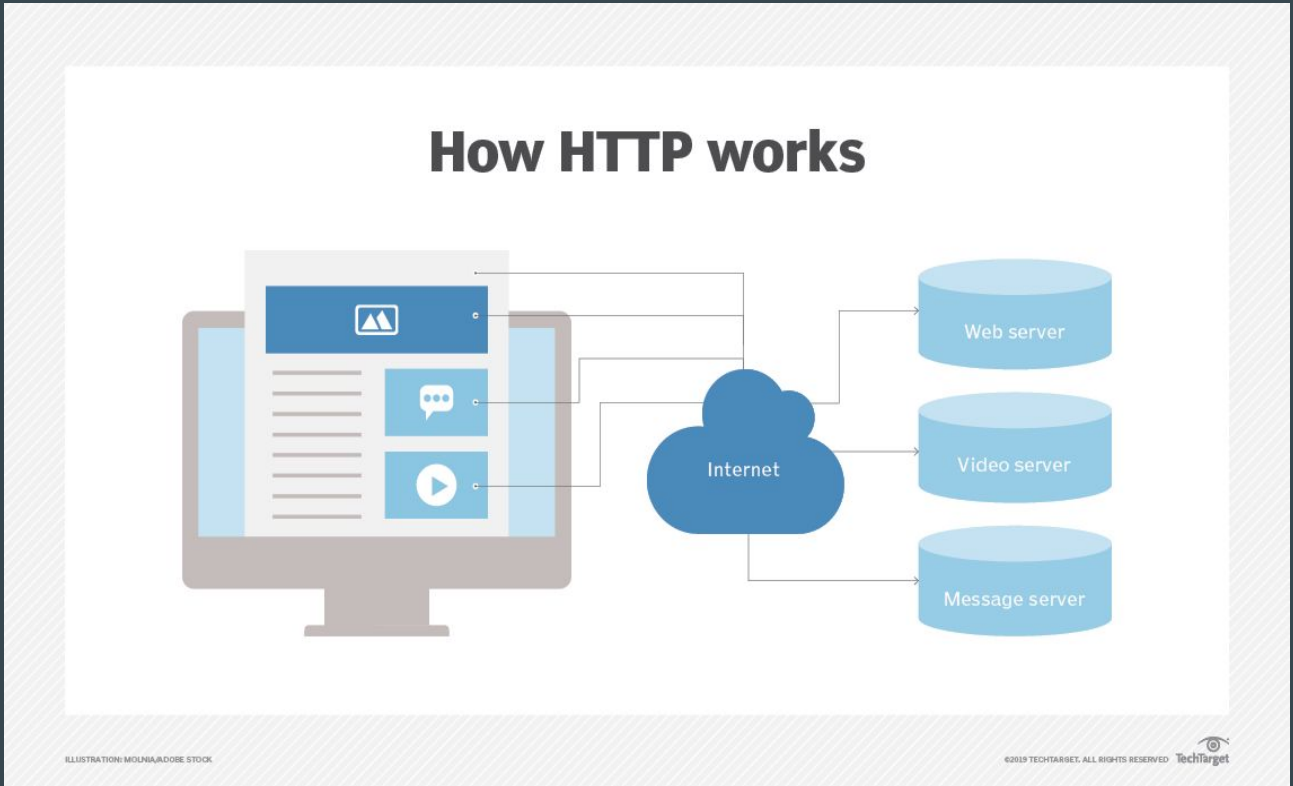
1. Overview of Relevant Web Technologies

1. Overview of Relevant Web Technologies

Relevant Web Technology:

- Hypertext Transfer Protocol (HTTP)
- Uniform Resource Identifier (URI)

Hypertext Transfer Protocol (HTTP)



- Request-Response protocol
- Client sends requests for resources.
- Servers send responses back to client.

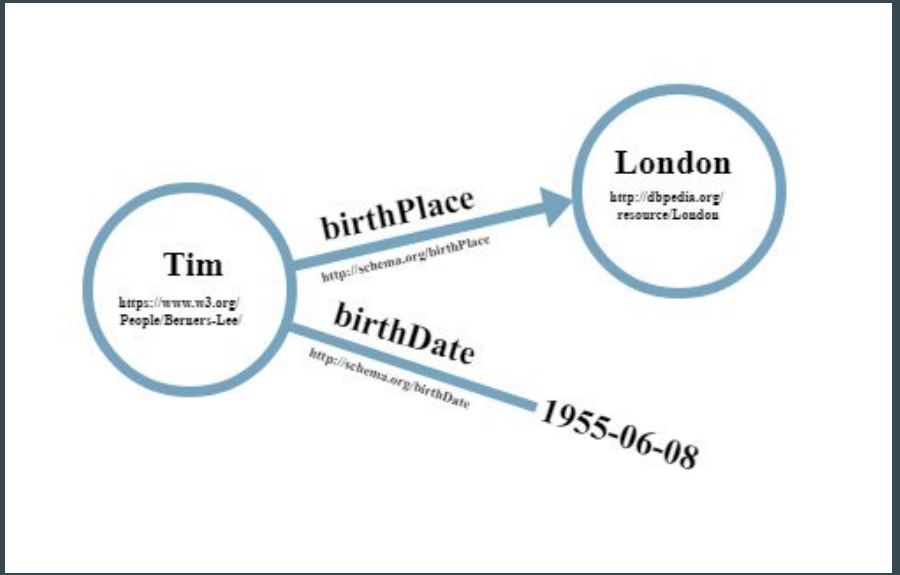
Uniform Resource Identifier (URI)

- Unique sequence of characters
- Describes resource
- Uniform Resource Locator (URL) is a specific type of identifier
    - Tells how to access resource
    - HTTP

2. Overview of Linked Data

2. Overview of Linked Data

- Set of Principles
  - Name starts with HTTP
  - Looking up an HTTP name returns useful data.
  - Anything else with a relationship also has name beginning with HTTP.



https://ontola.io/what-is-linked-data/

Linked Open Data

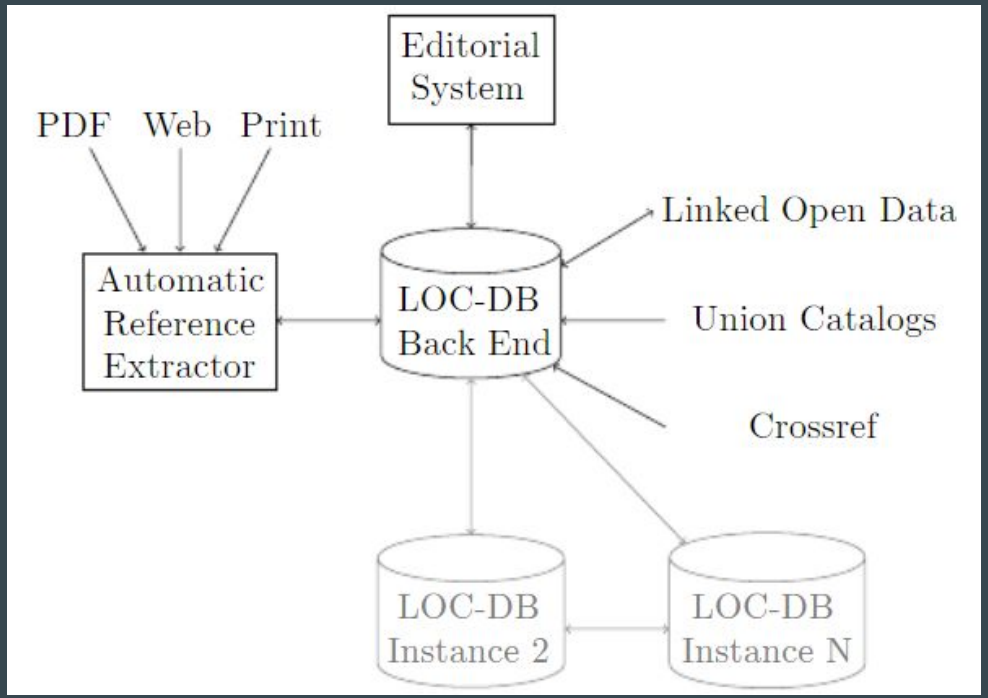
- Open data concepts
- Allows data to be accessible to any human or machine with access to the Internet.
- No paywall, copyright, or patent rights.

Applications of Linked Data for Library Use

Linked Open Citation Database (LOC-DB)

- Project created by researchers Anne Lauscher and colleagues.
- Current system:
  - Databases (DBs) of citations that are closed to other DBs.
- Goal: Create a Database of Citations
  - Designed for use in libraries
  - Citations taken from references within a work.
  - Linked Open Data Concepts
  - Designed to foster collaboration and sharing.

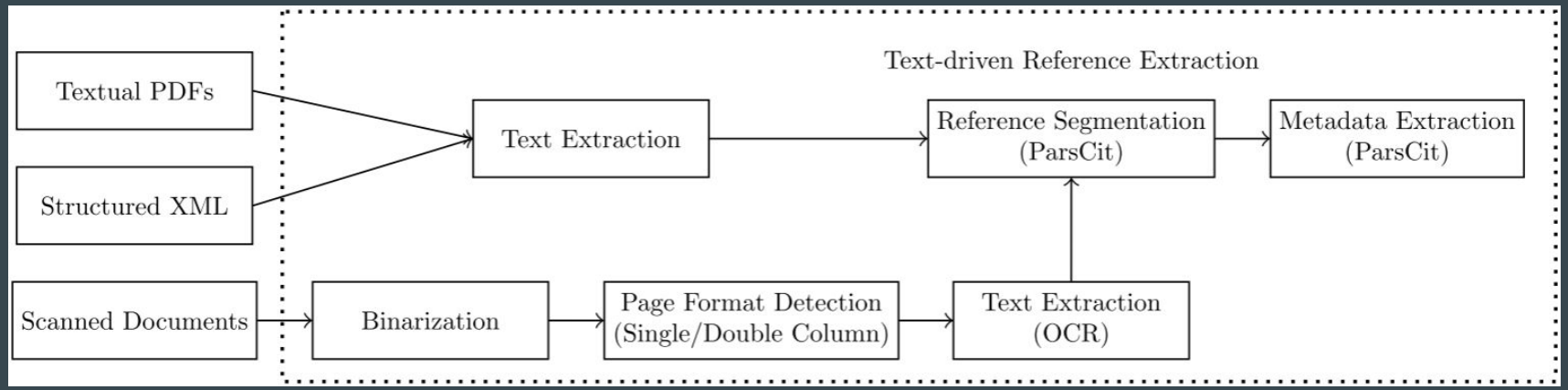
LOC-DB



- Semi-automated
- Distributed System
  - Separate Instances of DB
- Resource Metadata Creation:
  - Union Catalogues
  - External Suggestions
- Automatic Reference Extractor (ARE)
- Editorial System
  - Ingestion
  - Reference Linking
  - Metadata Editing

Source 1

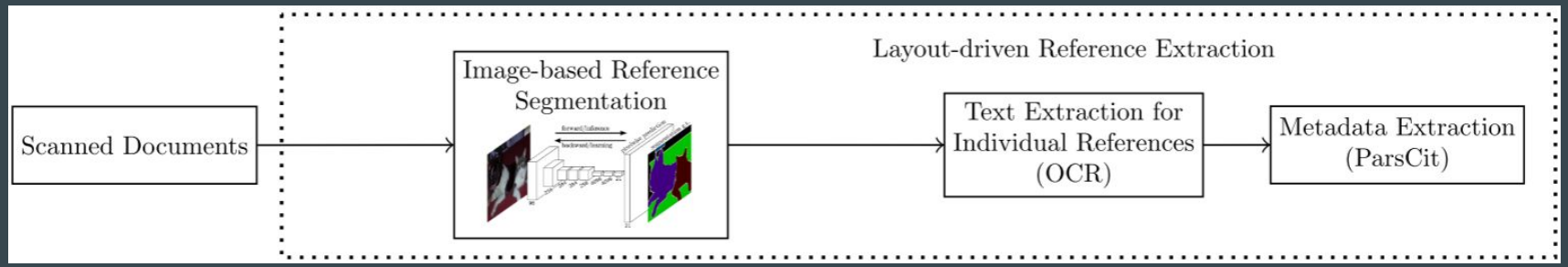
ARE Text-driven Reference Extraction



Source 1

- Optical character recognition (OCR)
  - Open source software OCRopus
- ParsCit
  - Open source package
  - Detect and label citation strings from text.

ARE Layout-driven Extraction



Source 1

Method	Total References Extracted	Extracted %age
Text-driven	3,645	71.7%
Layout-driven	4,323	84.9%

Source 1

User Study and Quantitative Evaluation

- Stakeholder Input:
  - Union Catalogues
  - Workflow issues
- Findings:
  - Scanning
  - Reference Linking

Criterion	Min	Max	Median
Citation Linking (s)	9.93	557.195	89.453
Internal Sug. Retrieval (s)	0.015	0.537	0.057
External Sug. Retrieval (s)	0.498	95.652	0.886
# Searches per Reference	1	36	2

Source 1: Minimum, maximum, and median time in seconds for the reference linking step with a sample size of 444.

Author's Take-aways

Pros:

- Open Data
- Both born digital and scanned documents
- Semi-automated component
  - Higher quality
  - Better results.

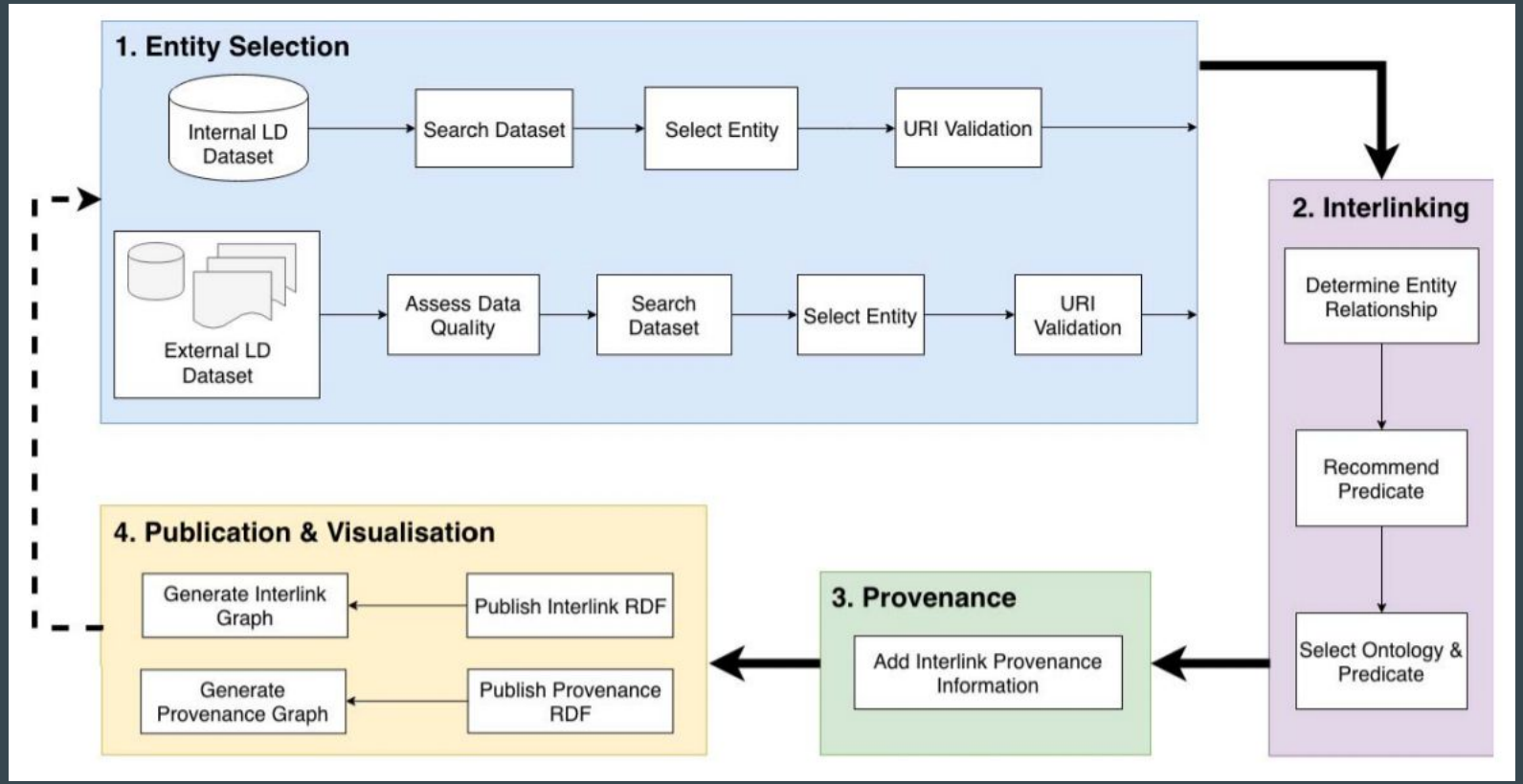
Cons:

- Current high cost to begin population of LOC-DB.
- No conclusive group to host LOC-DB.
- Many libraries needed.

The Novel Authoritative Interlinking of Schema and Concepts (NAISC)

- Researchers Lucy McKenna, Christophe Debruyne, Declan O'Sullivan
- Initial survey found:
  - Linked Data tools designed specifically for Information Professionals were desired.
- What they wanted:
  - Better tools designed specifically for libraries.
  - Work with the current workflow

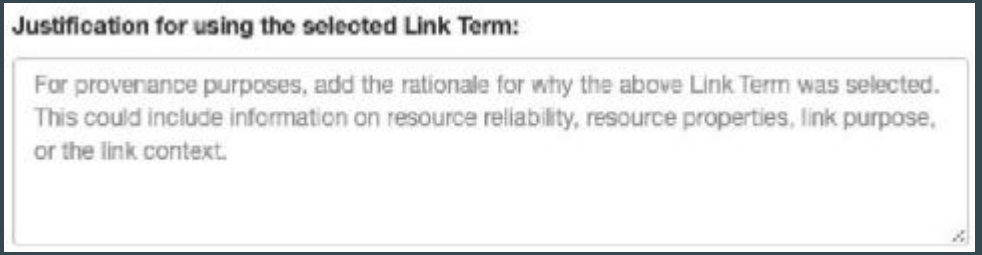
NAISC Interlinking Framework



Source 3

Provenance

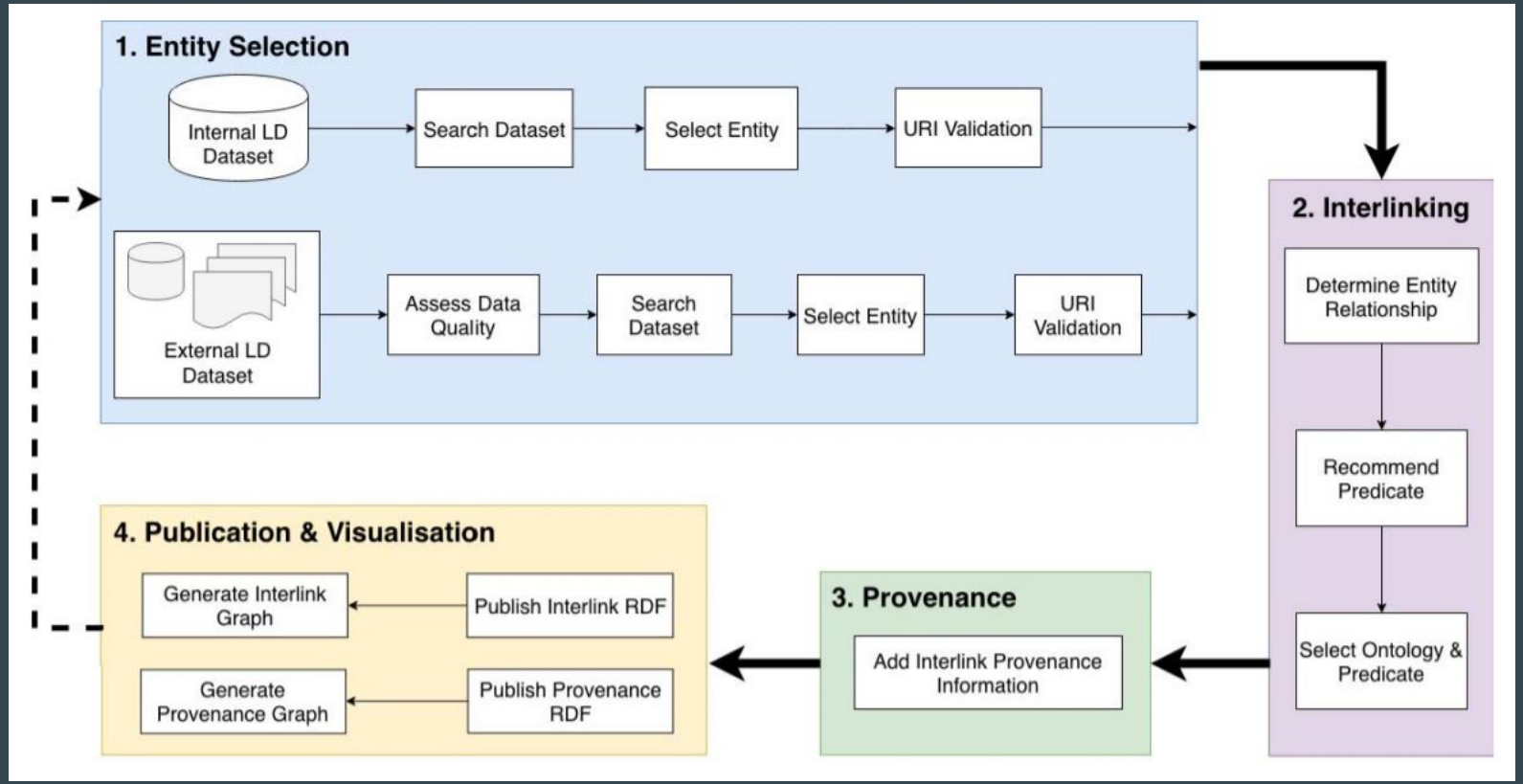
- Information about
  - People
  - Institutions
  - Resources
  - Processes



Source 3: NAISC Portion of GUI: Entity Interlinking

Source 3: NAISC GUI: Partial Provenance Graph

NAISC Interlinking Framework



Source 3

User Evaluation

P	Task 1		Task 2		Task 3		Task 4		Task 5		Task 6		Total	
	Complete (%)	Time (mins)	Complete (%)	Time (mins)	Complete (%)	Time (mins)	Complete (%)	Time (mins)	Complete (%)	Time (mins)	Complete (%)	Time (mins)	Complete (%)	Time (mins)
1	100%	1.766	100%	3.633	100%	2.433	100%	13.366	100%	2.683	100%	0.066	100%	23.947
2	100%	1.383	100%	3.75	100%	3.916	100%	12.133	100%	2.75	100%	0.133	100%	24.065
3	100%	0.866	100%	1.833	100%	1.7	100%	8.283	100%	2.016	100%	0.183	100%	14.881
4	100%	3.766	100%	4.0	100%	5.566	100%	33.566	100%	2.45	100%	0.366	100%	49.714
5	100%	1.5	100%	2.833	100%	3.633	100%	22.25	100%	2.8	100%	0.116	100%	33.132
6	100%	1.066	100%	2.483	100%	4.133	100%	14.2	100%	1.633	100%	0.05	100%	23.565
7	100%	0.75	100%	4.1	100%	3.816	100%	18.3	100%	0.85	100%	0.2	100%	28.016
8	100%	1.633	100%	4.00	100%	3.233	100%	28.033	100%	1.591	100%	0.2	100%	38.69
9	100%	0.866	100%	4.1	100%	4.55	100%	23.116	100%	4.13	100%	0.333	100%	37.095
10	100%	1.15	100%	3.3	100%	2.933	100%	11.283	100%	1.216	100%	0.45	100%	20.332
11	100%	3.9	100%	4.816	100%	4.133	100%	28.75	100%	3.883	100%	0.2	100%	45.682
12	100%	2.283	100%	3.966	100%	3.116	100%	18.05	100%	0.083	100%	0.25	100%	27.748
13	100%	1.6	100%	6.4	100%	4.1	66.66%	15.166	100%	1.033	100%	0.266	94.44%	28.565
14	100%	0.566	100%	3.35	100%	2.566	100%	14.6	100%	1.05	100%	0.233	100%	22.356
15	100%	1.383	100%	6.9	100%	4.583	66.66%	15.7	100%	1.525	100%	0.2	94.44%	30.291
Avg.	100%	1.631	100%	3.964	100%	3.627	95.55%	18.453	100%	1.983	100%	0.216	99.25%	29.874

Source 3: NAISC Think Aloud Evaluation

Author's Take-Aways

Pros:

- Interlink across datasets
- Facilitates increased engagement in Linked Data interlinking process.

Cons

- Current GUI
- Time consuming

Discussion

Acknowledgements

Thank you to Elena Machkasova for her feedback and guidance.

Questions

References

References

1. A. Lauscher, K. Eckert, L. Galke, A. Scherp, S. T. R.Rizvi, S. Ahmed, A. Dengel, P. Zumstein, and A. Klein. Linked open citation database: Enabling libraries to contribute to an open and interconnected citation graph. In Proceedings of the 18th ACM/IEEE on JointConference on Digital Libraries, JCDL '18, page109–118, 2018.
2. L. McKenna, C. Debruyne, and D. O'Sullivan. Understanding the position of information professionals with regards to linked data: A survey of libraries,archives and museums. In Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries,JCDL '18, page 7–16,, 2018.
3. L. McKenna, C. Debruyne, and D. O'Sullivan. NAISC:An authoritative linked data interlinking approach forthe library domain. In2019 ACM/IEEE JointConference on Digital Libraries (JCDL), pages 11–20, 2019.