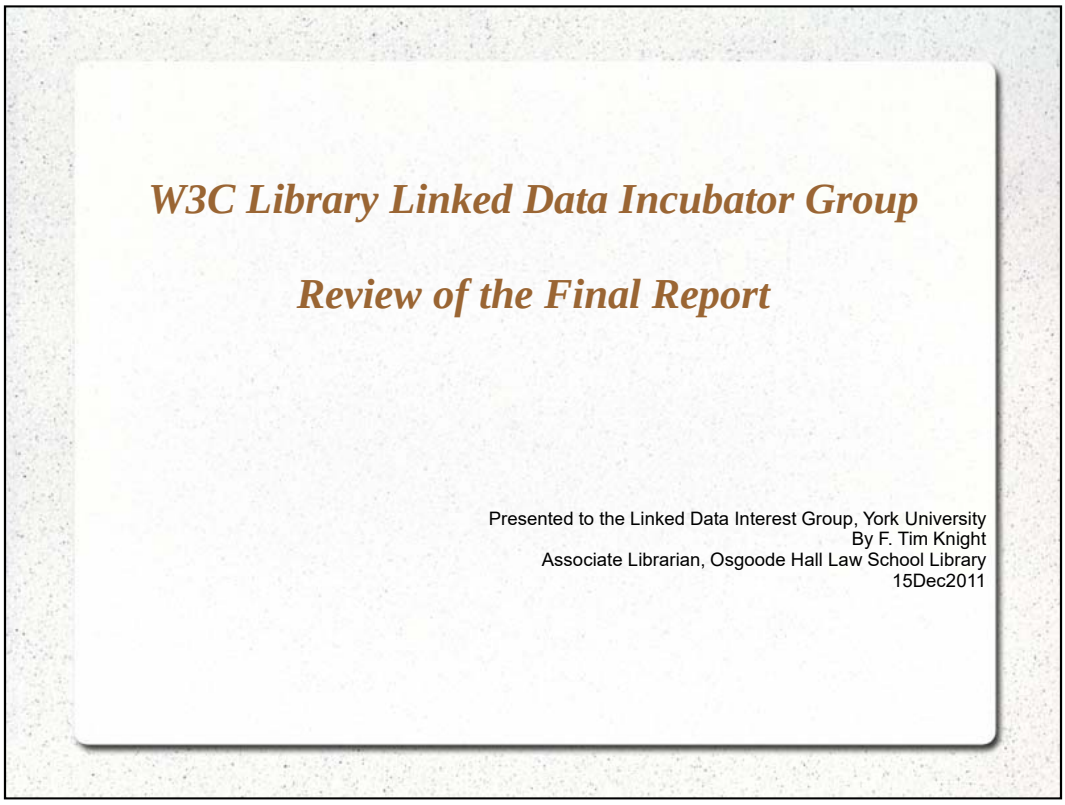
12-15-2011

W3C Library Linked Data Incubator Group: Review of the Final Report

F. Tim Knight
Osgoode Hall Law School of York University, tknight@osgoode.yorku.ca



This report is a snapshot describing the current state of library data management. It outlines the potential benefits of publishing library data as Linked Data and provides recommendations for library standards bodies, data and systems designers, librarians and archivists, and library leaders.

The authors represent international leaders in the library linked data field:

Thomas Baker, Dublin Core Metadata Initiative, US (W3C Invited Expert)
Emmanuelle Bermès, Centre Pompidou, France (W3C Invited Expert)
Karen Coyle, Consultant, US (W3C Invited Expert)
Gordon Dunsire, Consultant, UK (W3C Invited Expert)
Antoine Isaac, Europeana and Vrije Universiteit Amsterdam, Netherlands
Peter Murray, LYRASIS, US (W3C Invited Expert)
Michael Panzer, OCLC Online Computer Library Center, Inc., US
Jodi Schneider, DERI Galway at the National University of Ireland, Galway, Ireland
Ross Singer, Talis Group Ltd, UK
Ed Summers, Library of Congress, US
William Waites, University of Edinburgh (School of Informatics), UK

Jeff Young, OCLC Online Computer Library Center, Inc., US
Marcia Zeng, Kent State University, US (W3C Invited Expert)

There are two supplementary reports that provide additional detail. The first is the "**Use Cases**" describing library applications that take advantage of the benefits of adopting Linked Data standards and principles involved in publishing things like bibliographic data, concept schemes, and authority files. The second supplementary report "**Datasets, Value Vocabularies, and Metadata Element Sets**" provides a list of resources available for creating library Linked Data . There are several additional documents available on the W3C's **Semantic Web wiki** <http://www.w3.org/2001/sw/wiki/LLD> and there is discussion list **public-lld** <http://lists.w3.org/Archives/Public/public-lld/>, which are both open to interested members of the public.

Main Points of the Report

- Benefits of library linked data
- Current issues of traditional library data
- Library linked data initiatives
- Legal rights over library data
- Recommendations for next steps

2

Defining "Library linked data"

Terms Used in the Report

- Library
- Library data
- Linked data
- Open data
- Library linked data

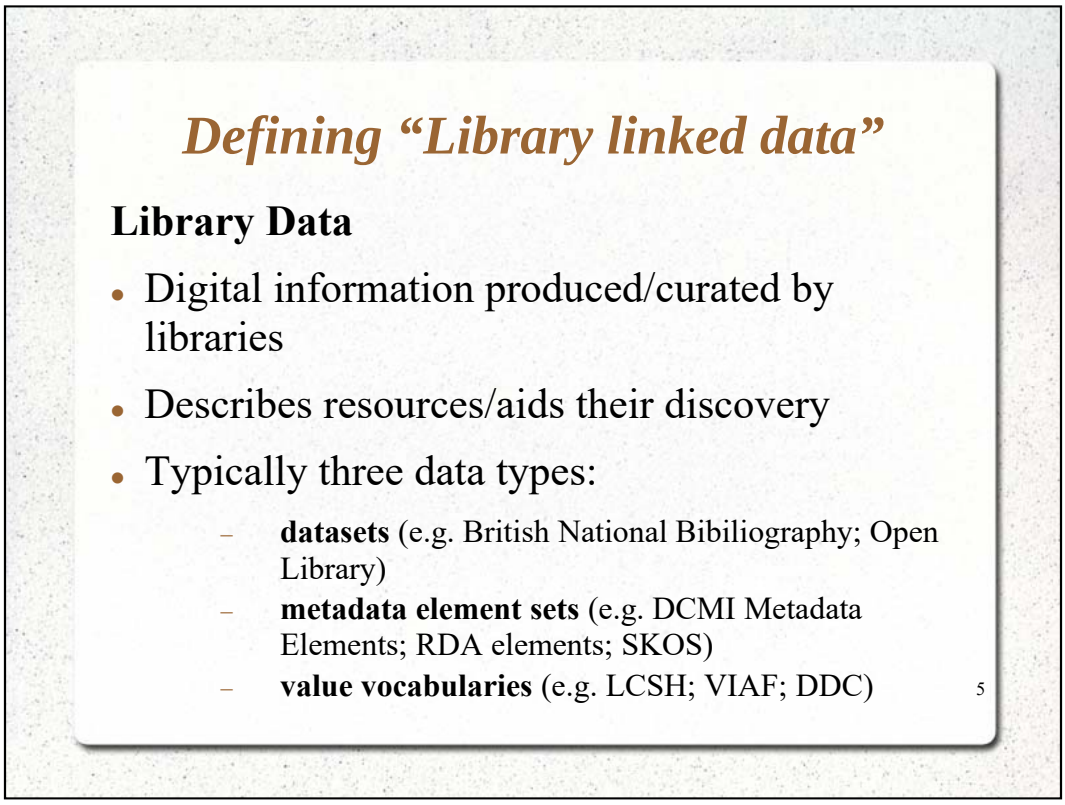
3

Defining "Library linked data"

Library

- Cultural heritage and memory institutions (includes libraries, museums, archives)
- A collection, a place, an agent (curator/administrator)

4



Datasets are data that reflect specific collections, e.g. British National Bibliography, Open Library, other national libraries, etc.

Element sets include things like the DCMI Metadata Elements; the RDA vocabulary elements; SKOS (Simple Knowledge Organization System); FOAF.

Value vocabularies are things like LCSH, VIAF (Virtual International Authority File); DDC; etc.

Defining "Library linked data"

Linked Data

- Data designed to facilitate links between:
  - datasets
  - metadata element sets
  - value vocabularies
- Expressed using standards (e.g. RDF)
- Defining relationships for navigation or integrating information from multiple sources

6

Defining "Library linked data"

Open Data

- "Legally interoperable"
- Freely usable, reusable and redistributable
- Licence requirements limited to "attribute" and "share alike"

7

Defining "Library linked data"

Library Linked Data

- Any type of Library Data expressed as linked data

8

Benefits of linked data

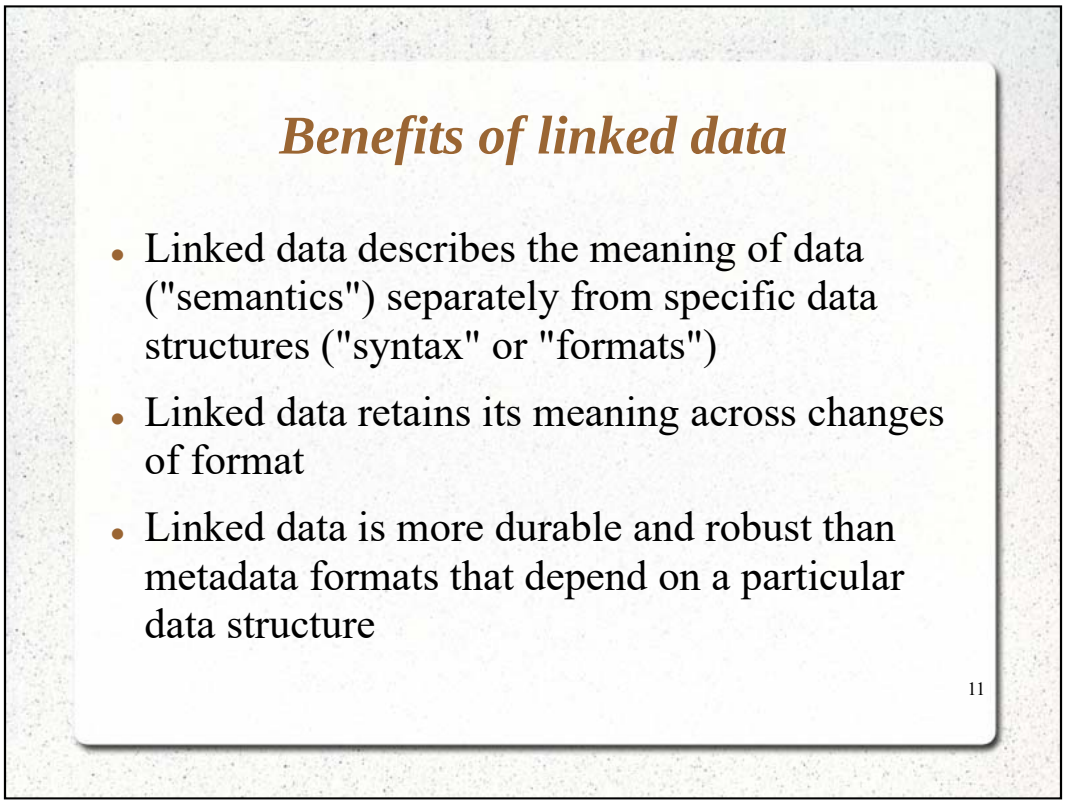
- Libraries can increase the value of their data beyond the individual sum of their sources
- Resources can be cited across a broader range of data sources
- Metadata descriptions become more accessible
- Capitalizes on the Library's ability to provide trusted metadata for resources of long-term cultural importance

9

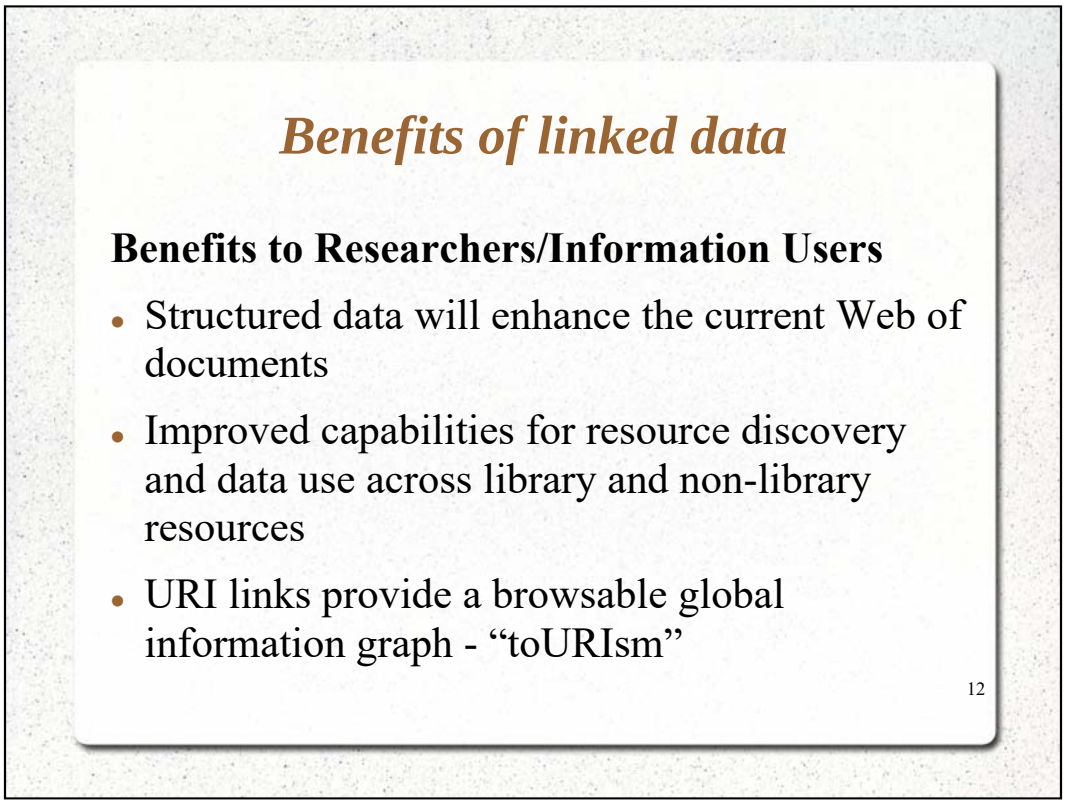
Benefits of linked data

- Data distributed as statements rather than complete records
- "Graph-based ecosystem"
  - Individual statements about resources
  - Collected and aggregated into a "global graph"
- Discovery of important connections made from previously unknown sources
- Redundancy of metadata descriptions reduced

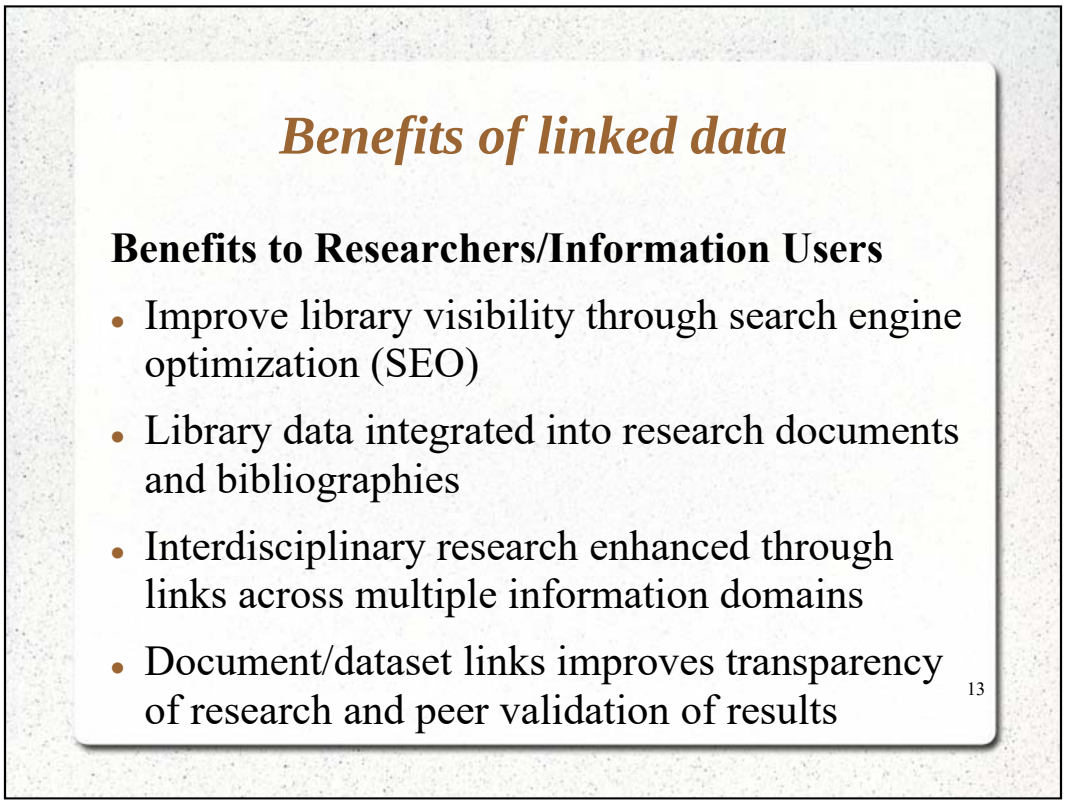
10



[Taken from 2.3 of the report and brought to an earlier position in this presentation]



The addition of structured data to the web should improve capabilities for resource discovery across both library and non-library resources (e.g. Wikipedia, news agencies like BBC or the New York Times, databases like MusicBrainz, etc.). It will provide users with a "richer set of pathways for browsing."

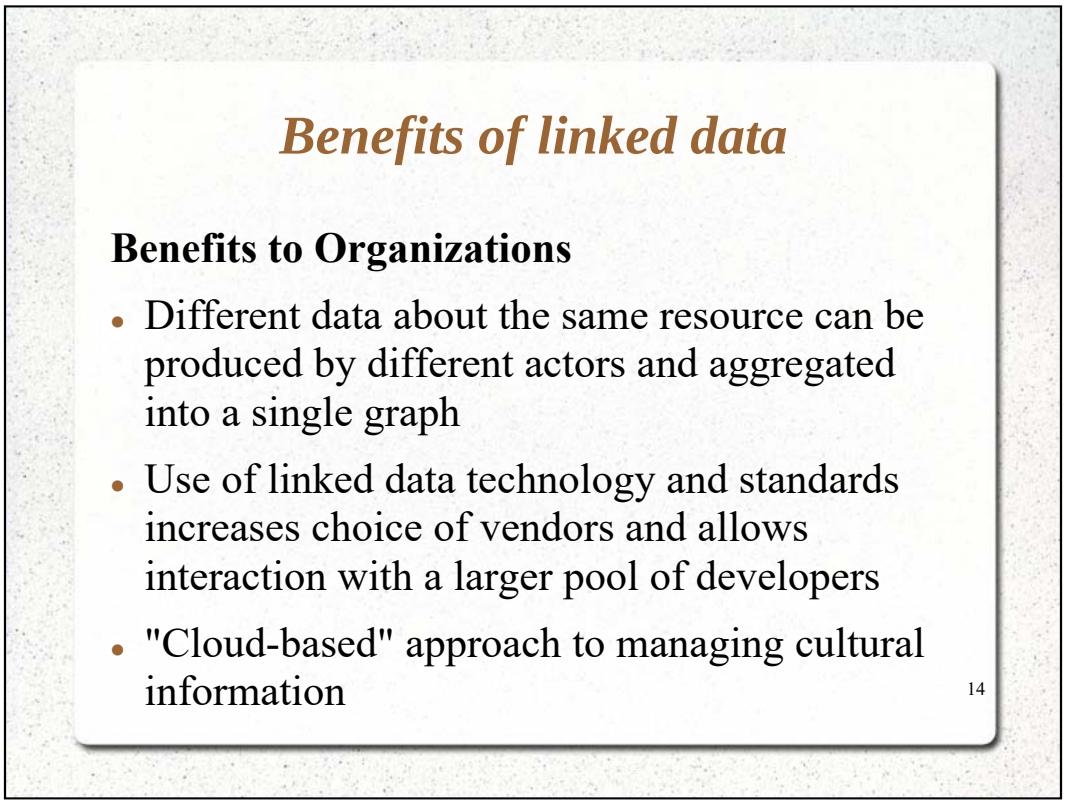


Search engines will make good use of URIs which will make library data more visible on the web.

Citation can become automated integrating library data into research documents and bibliographies.

Interdisciplinary research will improve as traditional barriers between subject domains are reduced and links made across multiple domains.

Providing links between research results and the data used to create the reports will make it easier to replicate the work done or reuse the datasets in different research models or for different purposes. This could make assessment of research and validation by peers easier.



Linked data is a bottom-up approach to publishing data that through aggregation of portions of descriptions can create aggregated descriptions with a much higher level of granularity then currently possible. The current top-down approach, creating complete descriptions in bibliographic records is labour intensive and likely unsustainable. Libraries have difficulty keeping up with their current workloads and are unable to achieve an increased level of granularity of descriptions.

Today's library technology is specific to library data formats and is provided by an Integrated Library System industry specific to libraries. If libraries adopt linked data technology it could give libraries a wider choice of vendors, and through the use of standard linked data formats allow libraries to recruit from and interact with a larger pool of developers.

Linked data could provide the first step toward a “cloud-based”/collaborative approach to managing cultural information which could be more cost-effective than stand-alone systems in institutions. This approach could also make it possible for small

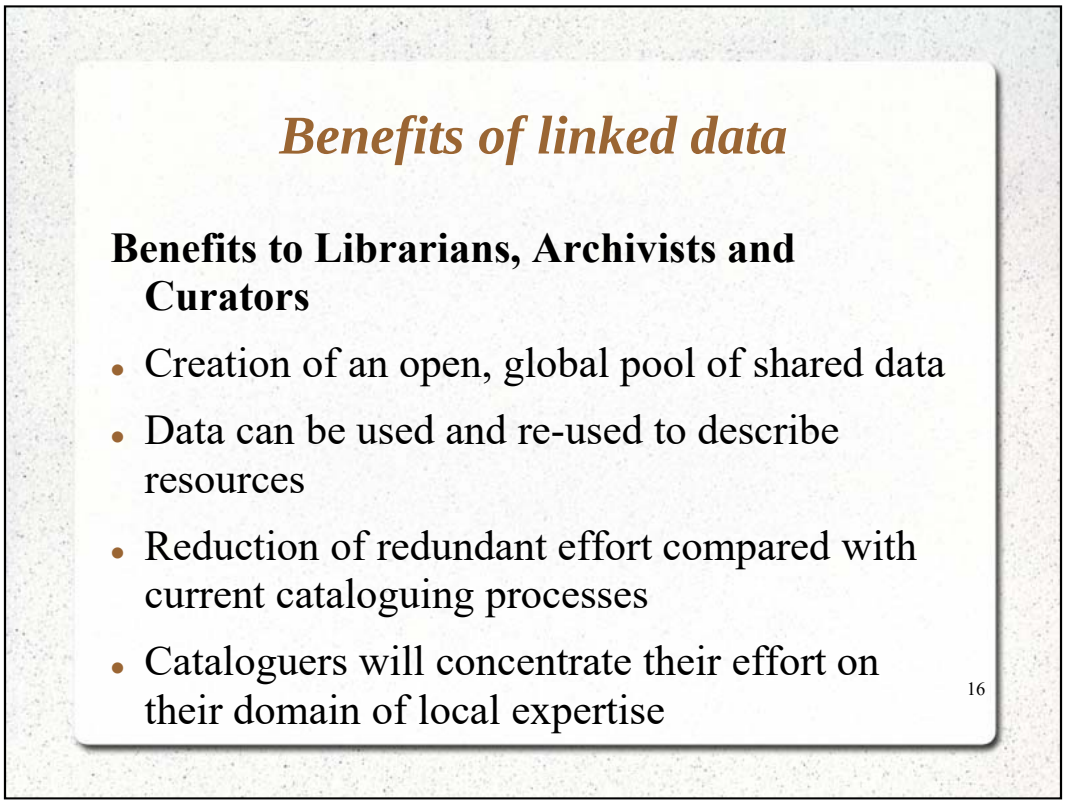
institutions or individual projects to make themselves more visible and connected while reducing infrastructure costs

Benefits of linked data

Benefits to Organizations

- Openness of data should be considered an opportunity rather than a threat
- Clarification of the licensing conditions of descriptive metadata facilitates reuse and improves institutional visibility
- “The coolest thing to do to your data will be thought of by someone else”

15



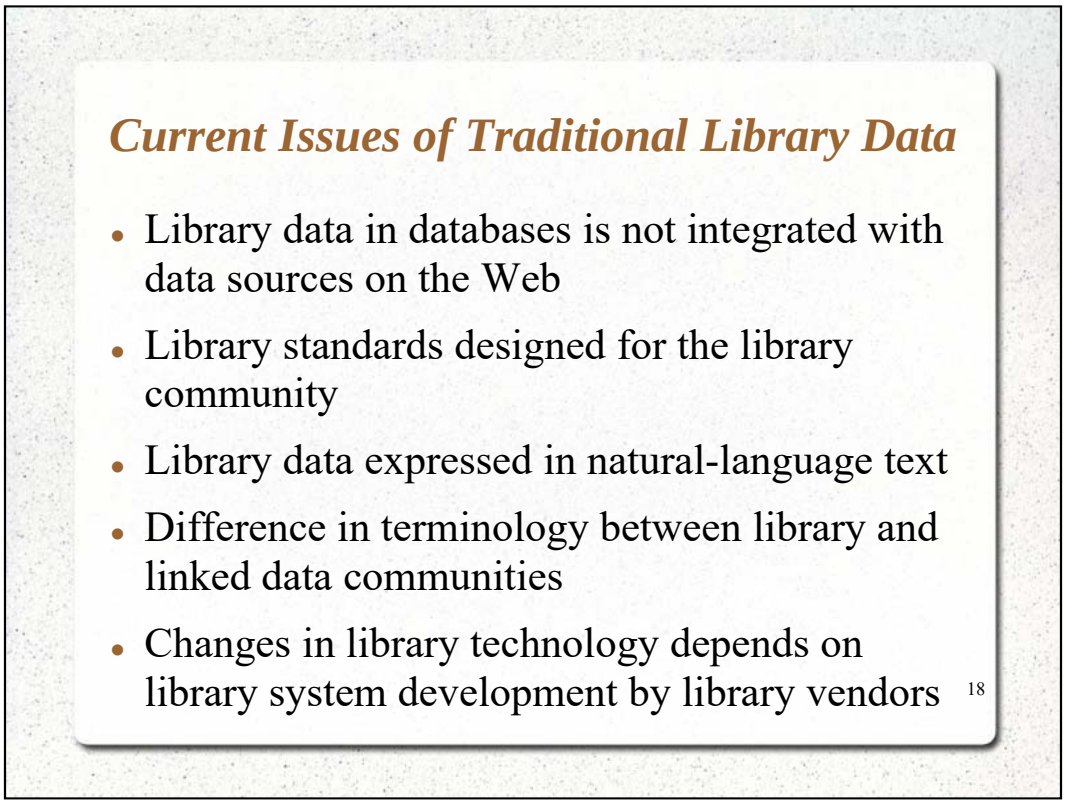
Provides cataloguers with more time for subject analysis and working with the intellectual content of the resource.

Benefits of linked data

Benefits to Developers and Vendors

- By leveraging RDF and HTTP development not tied to library centric formats and protocols, e.g. MARC, Z39.50
- Marketing products outside of library world
- Creates a larger developer community to support information technology in libraries

17



The main problem with library data is that it lives in databases that are not integrated with other data available on the Web. Theses catalogues have a web-based component but the data is independent.

This problem is attributable in part because the standards used in the library community were developed specifically by and for the library community, e.g. MARC and Z39.50. By using linked data standards this will broaden the usability of library standards with other data communities.

Library data is display oriented and meant for the most part to be consumed by the human reader. Even numeric values such as the ISBN is presented in a text field. The data is also managed locally and not globally. So for example a change in the OCLC database is not reflected in the databases of the contributing libraries. If URIs were used then the changes would be reflected in all places when changed in the central repository.

Each community has its own vocabulary, and these reflect

differences in their points of view, e.g. complete records vs. metadata statements.

Library Linked Data Initiatives

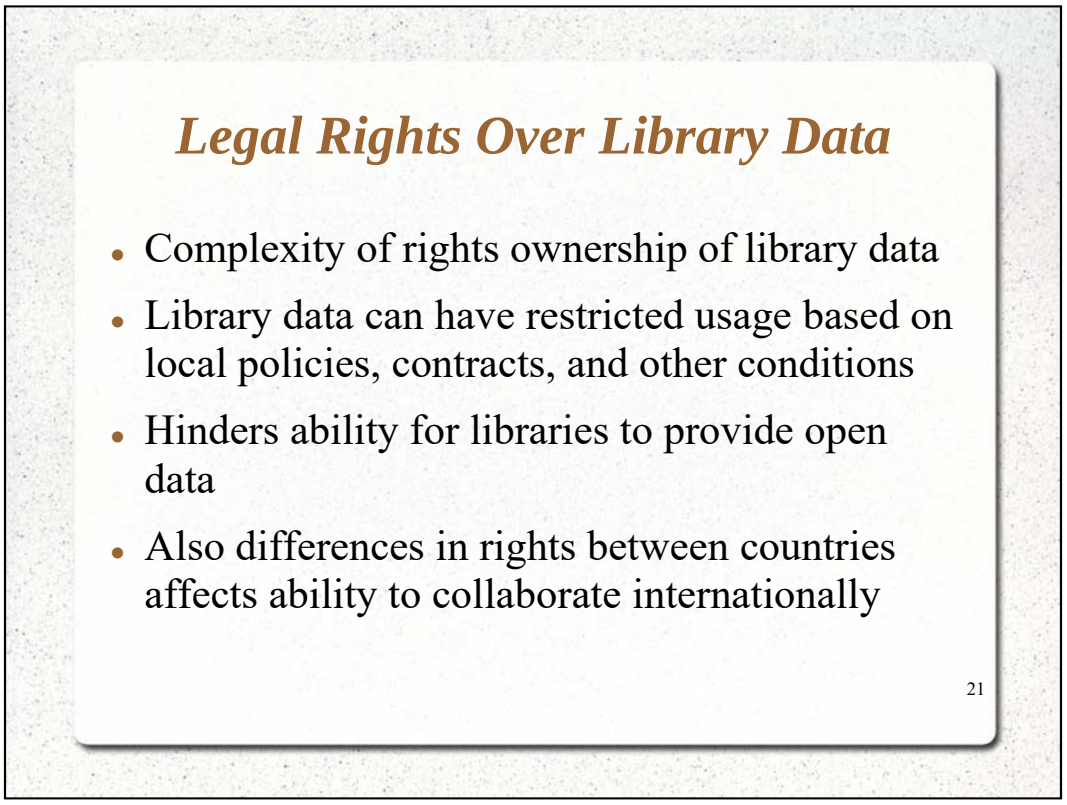
- Focus has been primarily on value vocabularies and metadata element sets rather than datasets
- Few bibliographic datasets made available
- Lacking metadata for journal articles, citations and circulation data
- Challenges: licensing; data modelling; dealing with legacy data; multiple user communities

19

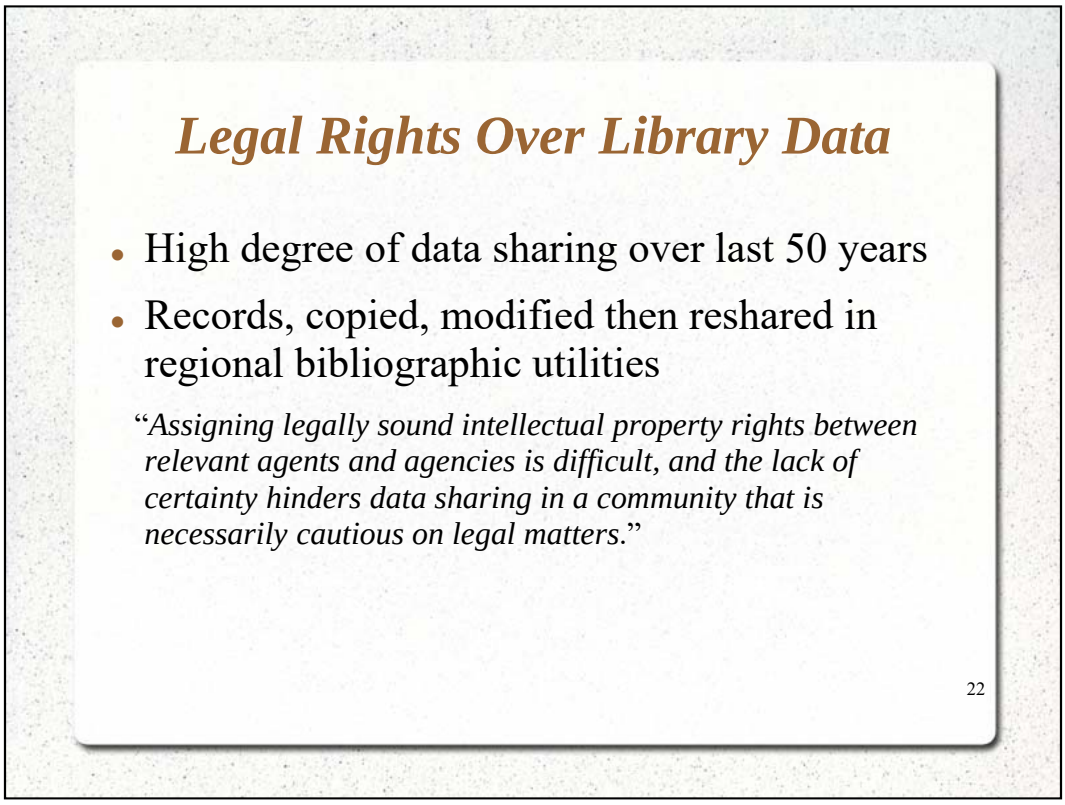
Library Linked Data Initiatives

- Increased interest and activity but largely "prototypes" so long term stability is unclear
- However many national libraries have started work in this area which will contribute to long term stability
- Need to develop connections across datasets
- Need for long term support; better communication between developers; mature tools to produce semantic links

20



One area that tends to hold the library community back.

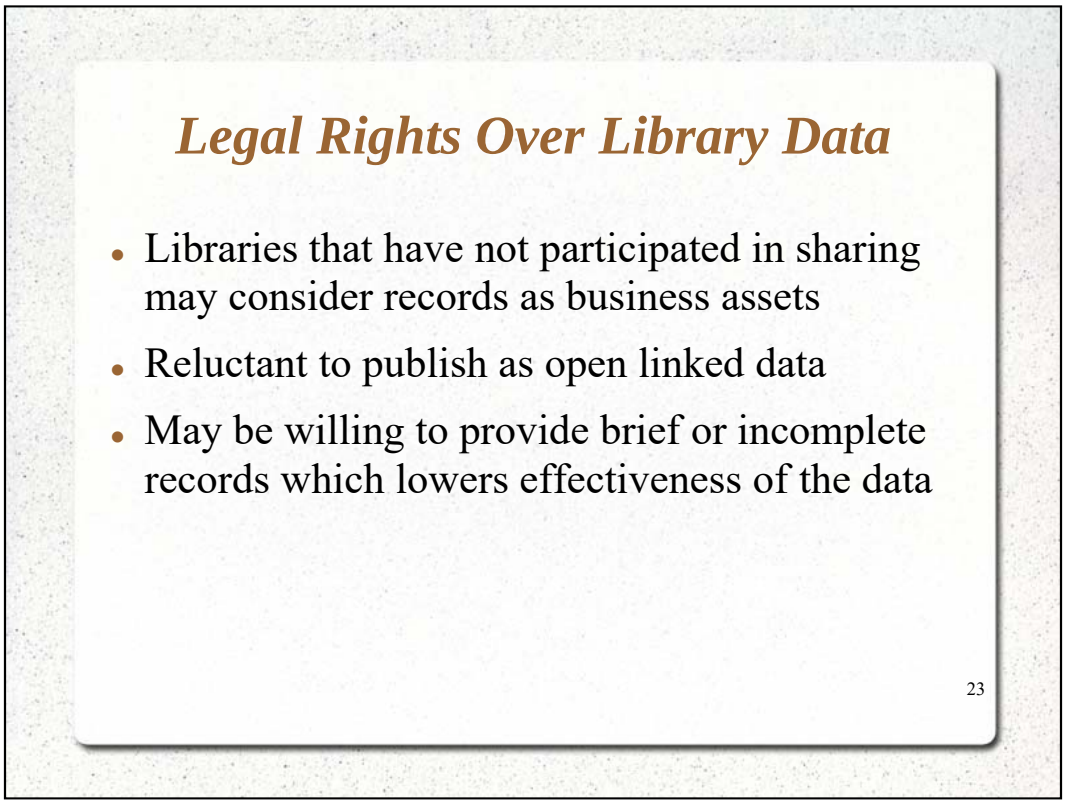


One of the strengths of the library profession is the general willingness to share bibliographic records. Libraries will look for copies of already completed cataloguing, copy the record and modify it for their local use.

When it comes to rights and ownership this can become a major weakness because it can be difficult to identify who owns the record. This is complicated by the fact that these records are often also uploaded to large bibliographic databases like OCLC who also want to lay claim to the content of these records.

This lack of certainty can hinder data sharing in an open way.

As noted in the Report: "Assigning ..."



On the other side of this coin there are libraries that have been working in isolation and consider their records as business assets. Something that they could sell to other libraries, for example.

There is therefore a reluctance to openly share the data they have created.

Some may be willing to share records that don't contain all metadata so-called “dumbed down” records or records that are brief, truncated, or otherwise incomplete.

Recommendations

- Generally if libraries make their data available for use as Linked Data it will integrate library data into web of information, provide greater visibility and bring library services to information seekers
- Libraries can lead by:
  - Managing resources for current use and long term preservation
  - Describing resources based on standardized rules
  - Responding to the needs of information seekers

24

Recommendations

For Library Leadership

- Identify sets of data as possible candidates for early exposure as Linked Data
  - identify high-priority, low-effort Linked Data projects, e.g. authority files and controlled vocabulary lists
- Foster discussions about Open Data and rights
  - seek agreement with owners about rights and licensing at the level of library consortia or even on a national or international scale

25

Recommendations

For Standards Bodies and Participants

- Increase library participation in Semantic Web standardization
- Develop library data standards that are compatible with Linked Data
- Develop and disseminate best-practice design patterns tailored to library Linked Data

26

Recommendations

For Data and Systems Designers

- Design and test user services based on Linked Data capabilities
- Create URIs for the items in library datasets
- Develop policies for managing Linked Data vocabularies and their URIs
- Express library data by re-using or mapping to existing Linked Data vocabularies

27

Recommendations

For Librarians and Archivists

- Preserve Linked Data element sets and value vocabularies
- Apply library experience in curation and long-term preservation to Linked Data datasets

28

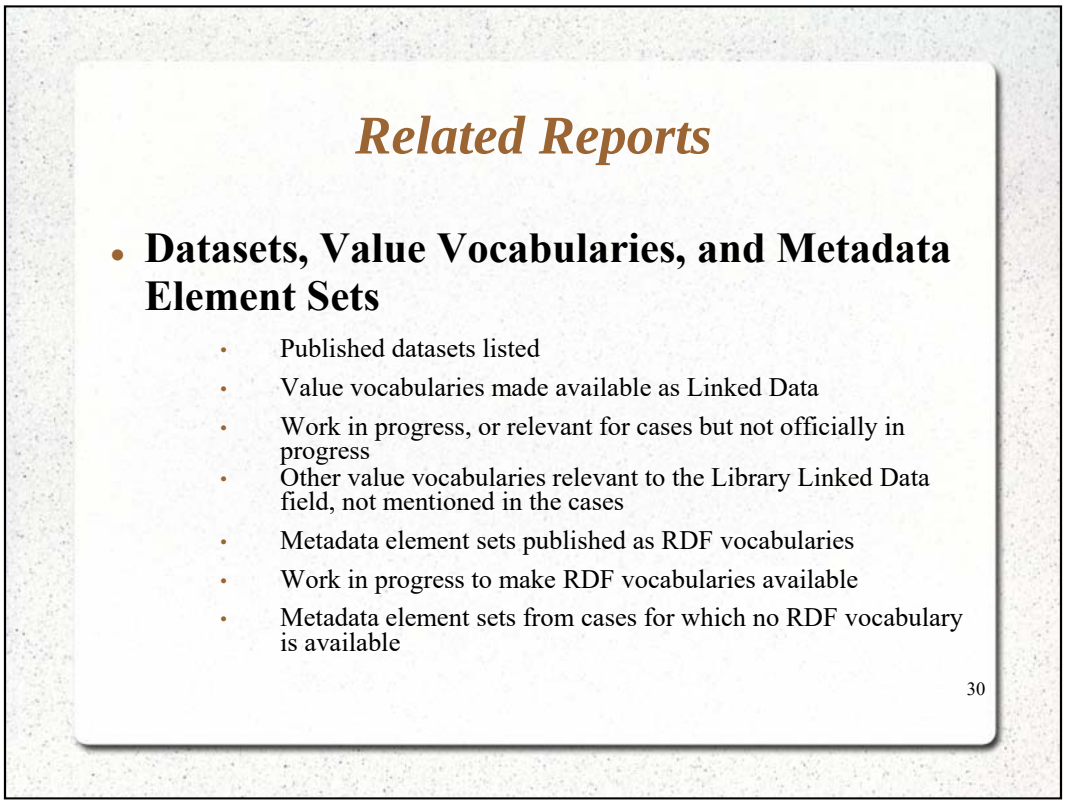
Related Reports

- **Use Cases**

8 “topical clusters”

1. Bibliographic data
2. Authority data
3. Vocabulary alignment
4. Archives and heterogeneous data
5. Citations
6. Digital objects
7. Collections
8. Social and new uses

29



Value vocabularies include: Classification systems; Subject headings/subject authority files; Name authority data; Thesauri; Other controlled vocabularies, e.g. DCMI Type Vocabulary.