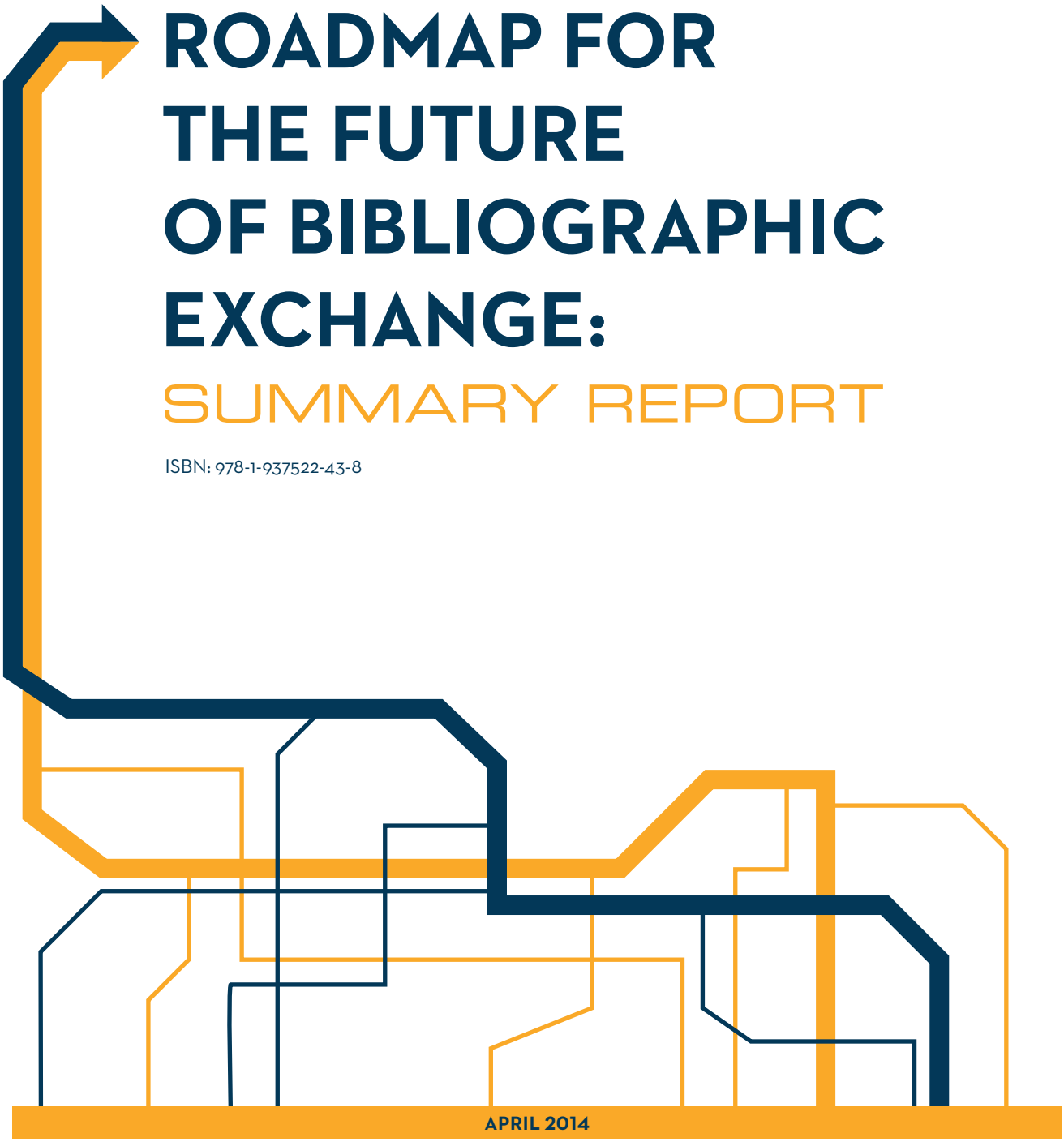
NISO
How the information world CONNECTS

ROADMAP FOR THE FUTURE OF BIBLIOGRAPHIC EXCHANGE:

SUMMARY REPORT

ISBN: 978-1-937522-43-8



APRIL 2014

3600 Clipper Mill Road, Suite 302, Baltimore, MD 21211-1948
P 301.654.2512 | F 410.685.5278 | nisohq@niso.org | www.niso.org

National Information Standards Organization

NISO's goal on this project was to determine the needs and requirements for extending the usability of the new bibliographic framework into the global networked information environment and to develop community consensus for a roadmap of activities needed in this space.

The vast majority of library users today prefer accessing information via the larger networked world, which demands approaches to data which can be more easily shared, indexed, and linked.

The background for this work, described in more detail in the grant proposal [http://bit.ly/1q8H8xj], illustrates a current landscape where most libraries are still creating and managing their extensive bibliographic data in MARC format. MARC, the *lingua franca* in libraries for over forty years, is often described as an outdated format, but its biggest liability in the modern web world is that it is unknown and unused outside of libraries. This uniqueness thus dooms library materials described with it to a siloed existence available within only library-oriented systems. The vast majority of library users today, who no longer consider libraries as the first point of entry for most of their information needs, prefer accessing information via the larger networked world, which demands approaches to data which can be more easily shared, indexed, and linked.

NISO's work in this area began with a two-day meeting in Baltimore in April 2013 attended in person and virtually by over 100 experts and participants, including librarians, system vendors, publishers, and consultants and vendors providing services around these. The meeting format, also openly available via a live stream, included lightning talks to highlight particular initiatives and provide some feel for how participants were trying to incorporate aspects of new bibliographic standards and sharing practices in their own environments, and then moved to a group brainstorming exercise where participants were invited to note particular problems that they thought should be addressed as part of libraries moving forward, or questions they had, or areas that should be discussed more broadly. Virtual participants were able to communicate their input via e-mail or Twitter. These notes were grouped into general themes, and then eight breakout groups were organized around the themes that gathered the most interest (from a brief show of hands). The theme-based breakout groups met for 60 minutes each, and used the notes as input for their discussions, but were not limited to these notes. The second day of the meeting included further lightning talks and then a full-group discussion of the overall themes, outputs, high level and low level concerns, and potential next steps.

Meeting Brainstorming Input Notes

Business Models

- Large sunk costs in existing MARC infrastructure
- Business model for creating high-quality subject/class metadata and metadata specialist positions
- Lack of use cases for linked data that libraries are willing to pay for
- Do we have to replace our system, reformat all our data, and retrain our professional staff?

Goals

- We need to be very clear and straightforward about what we're trying to achieve (better discoverability of library resources on the Web)
- Defined milestones to gauge progress for linked bibliographic data standards
- Encourage variety of systems with a shared vision
- Must work in existing environments
- Sustainability and maintenance of open bibliographic standards; need for an organizational framework
- Learning from MARC mistakes –What specific MARC-based problems to address in any system?

Interoperability

- Ontologies already in use in community
- What aspects of our data exchange can usefully be handled in common?
- How to publish mappings (e.g., existing mappings from ONIX to JATS/NLM)
- If everything is available via a URI, do we need to exchange anything?
- Enable end users to make use of data (e.g. mashups, scripting bibliographies) rather than forcing tools upon them

Openness/Sharing

- Make sure everyone is aware of open data frameworks
- Develop existing open data frameworks if needed
- Open source interfaces for data entry by metadata specialists that they can modify for their own use
- Understand the lifecycle of metadata in the connected world (i.e., local vs. global, distributed vs. centralized)
- How does an organization with lots of data share it through mappings? How to plug in?

Prototyping

- There needs to be a working prototype to demonstrate how linked data will work in libraries. We have to have something more than ideas to show
- Share code (and have a web clearinghouse of such code) for generating linked data from other kinds of data/metadata, e.g., entity extraction from text, etc.
- Inventory and encourage a variety of serializations with a same linked data vision.
- Flexible web apps that are not schema-dependent; for navigating concepts/resources –where are they?

Provenance/Authority

- Any record should be editable by any responsible authority with additions and changes validated by a central authority.
- How can we describe with sufficient provenance small, medium, and large data sets?
- How to hook all the author data (etc.) in the world to authority records?
- Provenance is a way for libraries to promote their trusted data.
- Is provenance simply "master record" in different form? Based on institutional reputation only?
- System integrity/security should be maintained through an authority/provenance system.

Users

- Empower users to create or contribute to bibliographic data.
- How to support researcher analysis of extractions from multiple databases (e.g., EBSCO, Gale, ProQuest) when APIs and content differ?

Moving Forward from Themes to Actions

Each breakout group was asked, after further exploring their breakout theme, to determine some action items that NISO could discuss further as part of the scope of this project. These action items were then collected into the NISO "Ideascale" idea-sharing website [http://bit.ly/1d10pap], where the project collaborators could then rank them and comment on them. More than forty potential action items representing worthwhile next steps came out of the meeting. The Ideascale tool was publicized in November 2013 and discussed on an open webinar in December 2013.

The two top-most ranked ideas from Ideascale were taken forward to an open discussion session held at ALA Midwinter, January 2014.

Action Items for Discussion on IdeaScale:

- Work to make vocabularies work across systems.
- Improve the ability of our data to be consumed and manipulated.
- Examine issues around authority.
- Describe use cases and business value/ROI for putting bib data on the Web.
- Create prototypes to determine format/interaction issues.
- Determine answers to questions around users.
- Further explore rules-related challenges.
- Create a forum and wiki space for sharing linked data projects.
- Address adoption barriers to new approaches.
- Understand impact of adoption on existing business models.
- Identify pain points in moving forward.
- Build interactivity; capture new awarenesses.
- Create partnerships with others in open spaces.
- Work on advocacy and education efforts.
- Define new business models and their value.

The two top-most ranked ideas from Ideascale were taken forward to an open discussion session held at ALA Midwinter, January 2014. These were: "Work to make vocabularies work across systems" and "Improve the ability of our data to be consumed and manipulated." Another idea ranked closely in third place, "Examine issues around authority," was an element of the discussion of the other two topics noted. The ALA Midwinter group then discussed the action ideas and determined some concrete projects, which could be implemented to help address the issues.

Make Vocabularies Work Across Systems

This action area initiated with the April breakout group discussing interoperability concerns as its theme. Some areas of potential effort that were named as part of that initial discussion were:

- Exploration of exchange strategies and application profiles
- Creation and management of semantic mappings
- Work on diversity issues, such as connection of bibliographic data to scientific data

Provenance of data is critical to understand the management needs of aggregated data as it ages and changes.

Many standards supporting interoperability are already in place; these include schema.org [https://schema.org], SKOS [http://www.w3.org/2004/02/skos/], RDFS [http://www.w3.org/TR/rdf-schema/], and OWL [http://www.w3.org/2001/sw/wiki/OWL]. Librarians who might use these in their own environments are not as aware of them as they could be, as they have been developed outside the library space by the World Wide Web Consortium (W3C) and other industry groups. Specific library-related application profiles and tools for libraries are needed.

Differences in vocabularies and the communities that manage them are often seen to be a hurdle to interoperability. Different vocabularies also present challenges because quality control, maintenance strategies, and usage policies vary across the sets. Provenance of data is critical to understand the management needs of aggregated data as it ages and changes. Quality checking often needs to be built into applications that use this data. The discussion group wondered, "Should vocabulary/data contributors be more explicit in communications about the quality assurance provided with their materials?"

Discovering existing vocabularies and associated descriptions and understanding the gaps between them can be a challenge, especially since many potentially useful vocabularies may not be oriented towards use in bibliographic data, but might have been created for geospatial, legal, or medical usage. In order to support improved awareness and discovery, a potential initiative could be to support the maintenance of a pool of available vocabularies and mappings, where individuals and institutions could potentially post assertions in order to show relationships. Each person or institution could map differently according to local requirements, but from this shared material a consensus could take shape. Some vocabulary users might also like to extend an existing vocabulary (once discovered/identified) to fit some local need or to reuse it in an alternative application. This possibility, though not technically complex, surfaces many policy and practice issues that need attention.

Questions arose about how specialized communities might be supported in this framework. Many expert vocabularies have evolved over time and have been extended to meet the needs of specific parties, and these provide a "world view" of a community of practice. This "world view" may be difficult to capture clearly in a semantic map, especially when the data is mapped across different knowledge domains (libraries, archives, museums). Other concerns include data loss and the need to retain original data during mapping between formats, where data elements and content differ. Approaches to these issues might be in the form of tools, training, or provision of examples and guidelines.

Different initiatives make their metadata available but few have documented their usage and expectations in ways that support interoperability.

Top-down efforts to impose order have been previously attempted by organizations such as IFLA, but it's generally agreed that there are limitations with reliance on only top-down approaches. Bottom-up approaches, perhaps encouraging a modicum of crowdsourcing power, should be encouraged—and a commitment to carrying provenance for all statements should allow users to determine for themselves what level of quality assurance is appropriate for their purposes, thus making many concerns about quality control moot. Another effective approach might be to support projects which allow different communities to expose their legacy vocabularies in standard ways, to support discovery and reuse by other communities, thus allowing extension, mapping, and maintenance to occur more broadly and cost effectively.

The discussion group at ALA in January 2014 discussed the problem from a slightly different direction: what type of metadata and amount of metadata is needed to discover an object on the Web (and implicitly, how can mapping vocabularies assist in the improvement of this process). To a great extent, the process of defining minimum sets of metadata must occur on a community or project level, preferably with explicit documentation as Application Profiles. Different initiatives make their metadata available but few have documented their usage and expectations in ways that support interoperability. Importantly, we must address the differences in requirements for needs other than discovery; e.g., for identification and disambiguation, or for long-term access/preservation purposes.

Some future activities proposed for NISO in this area include:

» Work specifically to bring related vocabulary efforts together to take better advantage of expertise, tools, and existing best practices. NISO's breadth of communities makes its leadership in this area invaluable. Organizing cross-industry groups to discuss potential use cases from different communities could be a good first step, along with an effort to publish recommendations on how existing standards can be applied across traditional membership groups.

- Explore existing stores of vocabulary information (the Linked Open Vocabularies project is a good start) to identify problems, gaps, and potential for collaboration. Given that collection and maintenance of vocabularies needs to include a global scope, which almost always includes language and translation issues, some coordination with IFLA could be useful.
- Ensure that NISO's own published vocabularies are in a machine-accessible form and take advantage of advancing knowledge in vocabulary expression and management. Recognizing that vocabularies published in human-readable PDF cannot provide the required functionality, NISO could seek to play a role as a model for organizations transitioning legacy standards to more modern representations.

In an environment emphasizing the principles of linked data, the notion of having to 'choose' a format for records is no longer relevant.

Improve the Ability of Our Data to be Consumed and Manipulated (by machines)

This potential action item emanated from the April breakout group which discussed the theme of "goals"—such as improving interoperability between existing systems and standards and enabling better visibility of library and cultural heritage data on the Web. To meet this goal, more attention must be paid to how data in this new environment could be more easily consumed, manipulated, and exposed to others by exploiting the power of machines. Libraries would be able to maximize innovation and minimize costs of a transition to a new framework if their data were accessible quickly and without barriers to reuse.

The ALA Midwinter discussions in this area covered the environmental or educational issues that arose from exploring the current landscape of linked data and sharing linked data. Creation and sharing of linked data in the first place is seen as a major step forward towards an environment where library data can become more machine actionable. In an environment emphasizing the principles of linked data, the notion of having to "choose" a format for records is no longer relevant. As the understanding and use of RDF triples to express bibliographic data matures, it is possible to relegate "records" to a place as input or output, and the discussion necessarily shifts to how existing data can be improved to support more active linkages. At the very least, libraries can begin today to emphasize a transition away from traditional bibliographic descriptions and use linked data concepts for identifiers in record creation and maintenance and tools such as VIAF and linked data-friendly expressions derived from LC name and subject authorities. Appropriate models for maintenance of such data, in particular versioning and change notification, are critical in supporting the continuing use of legacy data. Providing guidelines and support for best practices in this area is vitally important as demand for linked-data-friendly authority data continues to increase. In the use and maintenance of authorities, it

may also be possible to create methods for third parties (such as employers or library contributors) to validate links, as a model for long-term maintenance of data.

Some linked data stores already exist and a few tools are available, but many people don't know how to use them, or aren't aware of where they could apply them—thus these issues may be educational rather than technological in nature. The W3C Library Linked Data Incubator Group has gathered lists of tools and resources [http://www.w3.org/2001/sw/wiki/LLDtools] and those lists could be updated. Several of the large players, among them the British Library and OCLC, have exposed some of their data as linked data. It's unclear how the majority of libraries can make use of this data or the incredible variety of data stores available on the open web, for instance those created and maintained by Wikipedia and the New York Times. It might be useful to communicate with those who are working on prototypes to get their feedback on what directions they see as needing attention.

Rights are often an area where innovation meets uncertainty.

In order for library data to be effectively shared and manipulated, libraries must become further educated on the "power of the statement" versus the more traditional "record." Librarians used to working with full MARC records may not easily grasp that a move to the more atomic level of individual statements will make possible innovation in areas like new services, localization, and distributed data improvement. Outside of libraries, these activities are building and taking shape, but most librarians aren't yet monitoring those activities, mostly because they have yet to appreciate the connection with the library world.

It is desirable that data, at the point of creation, include unique identifiers supporting stronger links within and among systems. For example, repositories or publisher submissions systems could build into their submission processes requirements for inclusion of standard identifiers for the authors (e.g., ORCID or ISNI). The presence of these IDs could make linking of authors to their work more reliable and no longer subject to ambiguity of names (particularly in the sciences, where forename initials are still the norm). The organization sponsoring a project could also have an organizational identifier included, creating further opportunities to assist users in navigating the myriad of information sources available to them.

However, potential barriers to sharing data exist throughout the current environment. Linked data is often referred to as "linked open data" but the notion of "opening" their data conflicts with the business plans of some potential contributors, such as corporate entities who may see their data as more of a proprietary asset, even if they receive attribution. Rights are often an area where innovation meets uncertainty. For instance, data issued with CC BY licenses may or may not include potential

for commercial use (whether the license includes the –NC suffix), and some licenses defined for documents don't work well with data. A concern was expressed that changes that may be made in the course of reusing the data may affect the original publisher's authoritativeness and its "brand."

Data publishers who do not provide rights for these activities to consumers may see their data go unused and thus their contribution will not represent the advance they intended.

Other concerns voiced by entities with potential to contribute existing data sets included: publishing proprietary authority files would include data that was not 100% clean, potentially embarrassing to a corporate entity; there may be exposure or accountability (if not liability) issues if data that was made available for use in other applications somehow harmed or damaged a recipient application; and corporate entities may wish to have fully established business models for their data contribution before publishing their data. National libraries may see issues around the application of licenses for metadata and licenses for the objects themselves. There may be undesirable applications (e.g., criminal, exploitative) where potential contributors don't want to see their data used. Enforcement of CC licenses, even CC BY, may be seen as expensive or impossible.

On the side of linked data consumers, discussion explored whether an explicit CC0 license needs to be attached to data in order to make it attractive to potential users. CC BY-NC is an alternative license, but it is often unclear to users what "non-commercial" use actually means. It was observed that if potential consumers of linked data see restrictions on what they can do, the data won't be used or republished or improved by other users. Data publishers who do not provide rights for these activities to consumers may see their data go unused and thus their contribution will not represent the advance they intended. Licenses designed expressly for data such as the OpenDataCommons licenses [http://opendatacommons.org/licenses/] are now becoming available, but their application requires further education.

Many legal questions, not easily taken on by lay users, may arise and become barriers for easy adoption of linked data sets. For instance, if a consumer changes data and republishes it, does that change ownership of the data? Is ownership even a useful idea where data is exposed openly? What are best practices for attribution regarding authorship of data?

In addition, those who make investments in utilizing linked data need assurances that the data will persist and be updated. Thus all participants in a linked data environment need to understand the responsibilities behind their contributions and their reuse activities.

Some future activities proposed for NISO in this area include:

- Create a recommended practice or an informational document around the use of linked data and associated rights and their implications. Emerging linked data communities (such as those including national libraries and entities such as DPLA and Europeana) can contribute their experiences and perspectives. Though there still may be further understanding needed across the broader community on what you can and can't do with open data, this work could start the conversation.
- Create a community recommended practice specifically for data contribution for corporate entities to utilize as a justification for their contributions and potentially to use as a shield, or partial shield, in regard to liability questions.
- Organize, evangelize, and manage an authority file as an additional/alternative Registration Agency for ISNI to expose the ISNI to communities not familiar with the standard. The existing Registration Agencies for the ISNI standard may be seen by potential contributors as being too "corporate" thus a barrier to contribution of existing organizational data. (Pockets of this data still exist in a non-centralized way, signaling a need for a collaborative approach to remove barriers.)

Conclusion

The activities that were determined, through community discussion, to be part of the NISO Bibliographic Roadmap in large part aim to be applied to existing efforts and maximize their usability as much as possible. It was recognized in many discussions that though the larger library community overall may seem to be hesitant in moving forward amid a fair amount of uncertainty in the lack of a solid technical framework, there is already much experimentation and many projects under way in diverse spaces. Further practical exploration of existing vocabularies, linked data tools, and methods for data contribution can help to reassure the community that the transition forward will not be endless and the value of what libraries already do will be enhanced.

NISO, as a non-commercial industry organization based around the development of standards and recommended practices, intends to capitalize on its ability to attract disparate stakeholders together to increase interoperability in future efforts carrying this work forward. NISO's leadership, via the Content and Collections Management Topic Committee [http://www.niso.org/topics/ccm/] plans to examine these prioritized Roadmap work items—as well as the other ideas generated throughout this process—for future action during 2014 and 2015. Projects, which might take different forms depending on requirements and intended outputs, would be subject to the approval of the NISO Voting Membership. ■

Further Information:

NISO Bibliographic Roadmap Project Page includes meeting agendas, presentations, recordings: http://www.niso.org/topics/tl/BibliographicRoadmap/

Brainstorming Meeting Output/Notes: https://sites.google.com/site/nisobibrm/

IdeaScale Input Forum: Bibliographic Roadmap: http://bit.ly/1d10pap