Background Paper

Libraries and other cultural heritage and educational agencies create name registries or “authorities” to serve a variety of purposes, usually within an institutional or disciplinary context, but the data they create have potential for broader reuse. The April 2015 IMLS National Digital Platform Forum report emphasized the importance of enabling technologies (e.g., interoperability via linked data) and radical collaborations in supporting the mission of the cultural heritage sector. Our forum aims to bring together stakeholders working on shared local authorities in order to develop a shared understanding of the key issues. One of the goals of the project is to articulate those issues in a white paper that can serve as the basis for future work. But the meeting also aims to explore a range of possible avenues and outcomes to further the objective of making authorities more shareable.

In recruiting participants for this meeting we sought to include stakeholders representing a wide range of disciplinary and institutional perspectives. The common goal is to share the data we produce; but for this to become possible, it is necessary first to survey the variety of use cases and assumptions in each domain, and to understand preconditions for improving shareability. The following points emerged in a series of discussions held with group members leading up to the meeting, and are documented here as background for the October meeting. In considering these issues, participants may wish to consider the following questions:

- *Does this issue affect you? In what way?*
- *Who else is affected?*
- *What solutions look promising? Who is working on them?*
- *What kinds of joint action would help?*

Workflows vary widely. Library authority practice is based on creation of data by third party producers, but researcher identifier systems like ORCID typically rely on a self-registration process. Creation of data may be strictly local, or it may be distributed but on a highly controlled central platform, as in NACO. Ingestion or production processes may be set up to pre-empt duplication, or be relatively tolerant of it. If the latter, downstream resolution or remediation may become correspondingly more important. Data from any of these sources may be aggregated into a central hub like ISNI; there may or may not be a quality threshhold for publication of the data. Sharing of data of necessity raises the issue of propagating changes and keeping data in synch, and at a higher level also of understanding the dependencies among different services. As always, there will be trade-offs between workflows that are considered optimal and what the available infrastructure will support.

Data models will also need to be addressed to facilitate sharing. Differing treatments of pseudonyms, for example, can lead to conflicting representations of entities. At a more basic level, the vocabularies used to express these relationships also need to align if data are to be reused effectively across domains: for example, a hierarchical relationship must be understood to be such even when it crosses silos, since in addition to its value as information for end users it also enables contextualization. The need to support reuse in itself introduces the need for new relationships, such as explicit declarations of equivalence or indeed of non-equivalence. This is an area where linked data solutions can pay great dividends, but where reconceiving legacy practices in linked data terms can pose challenges that different communities are at different stages of meeting. These issues arise not only for newly created data, but even more acutely for the large bodies of legacy data still in need of reconciliation. Practical considerations make themselves felt here too, and the models we adopt may have to accommodate themselves to the limitations of legacy systems, at least for the time being. In addition, communities will have to navigate the policy issues that arise in a linked environment: for example, the use of alternative vocabularies, the relative priority given to widely established and heavily linked identities as against unique, local, or siloed ones, or the question of how to lower barriers to contribution if the desire exists to cast a wider net for identities. The diversity of strictly local practices reflecting historical contingencies - staffing, local interests, the quirks of organizational structure or legacy systems - pose their own obstacles to shareability.

Persistence is an issue that arises repeatedly in discussions about reuse. Best practices commonly identified include non-reassignment of identifiers, tombstoning of deprecated identifiers, and retention of selected data. These prescriptions are readily implemented in some areas, such as traditional library authorities, but present a greater hurdle to accustomed practices in others. In an ecosystem where data are freely traded **provenance** and **trust** also loom as fundamental issues. Modelling and best practice issues arise here too: in a distributed environment, provenance has to be captured with greater specificity than was necessary in a centralized, record-based ecosystem.

All of these issues have a **technology** dimension as well as a social and institutional one. Trust may be reflected in policies about preferred sources, but systems must then be able to enact those preferences, and manage distribution of updates to various partners. One of the clearer areas of emerging need is the ability for vocabularies to support services providing external lookups to external editing and reconciliation tools. As use of these **services** increases scalability becomes critical. So does sustainability, which is a precondition for persistence in identifiers, and in turn brings into play a range of institutional questions such as storage, documentation, and funding. Available tools often reflect the needs of the purpose they originally served and the workflows they were embedded in, and the need arises to evaluate their suitability against a broader range of uses. This also is information to be documented and shared.

Differing **institutional mandates** may be reflected in divergent practices. Organizations as various as national libraries, universities, scholarly publishers, value added service providers and disciplinary societies will have different imperatives and their use cases will show different emphases. These differences will make themselves felt in crucial areas such as the **business models** that sustain their services: sources of funding, such as grants or memberships, that sustain one area of activity may not be readily available in another where the value proposition is different. For example, a university may provide support for a researcher identifier system to track the scholarly output of currently affiliated faculty, but it is less clear if the same justification extends to past scholars or to the broader scholarly community. Yet shared data must of necessity be able to meet a range of uses outside its original ambit. **Licensing** of metadata is a difficult issue for related reasons. While unrestricted licensing of data is often cited as an ideal, it may be in conflict with the business model that supports the service in the first place, and some compromise may need to be struck. A related issue is **confidentiality** of data, where the effects of aggregation and reuse can have consequences not foreseen by the original producers of the data - or the people described by it.

Communities will have to address **governance** to respond effectively to needs in all of these areas, from compatibility of data models to scalable infrastructure to sustainable business models. They also need to articulate their goals, and their place in landscape they operate in, to their audiences. Because these are areas of shared need among participants in this forum, they are also potential areas for collaboration.