

Linked data for the stewardship of research knowledge in shared library, archive and museum (LAM) domains

Executive summary

Libraries, archives and museums (LAM) hold a vast quantity of data and resources that, until now, have often remained hidden from sight in catalogues and archives. Unlocking the potential to exploit and diffuse this previously unknown information to a wider audience would bring opportunities to further enrich the World Wide Web, promote a culture of openness towards knowledge and create a number of advantages for each individual link in the information chain. For libraries, archives and museums, this means a valuable opportunity to create bridges across collections and to provide more comprehensive tools to patrons, thus providing end users with a wealth of data, while data publishers are able to promote new forms of cooperation. Thus, the world's cultural heritage may be carried forward for the benefit of future generations, research may be preserved in its original languages, and cultural variety and vitality may be kept alive. The potential benefits to the entire Humanities sector and the many domains and niches that it encompasses which now more than ever risk marginalization, are significant.

Linked data is the instrument that can make this possible.

This new approach to data and information processing is laying the groundwork for further improvements in how institutions engage with each other, and for the reuse of data in varying contexts, enabling more efficient web-based entity identification and discoverability. At the same time, the new approach supports the use of cultural heritage collections, and encourages data to be used and shared more widely. By applying the linked data paradigm, libraries, archives, museums and information professionals will have a more comprehensive suite of tools at their disposal. These tools will evolve as institutions engage with each other in a common workspace and will have the freedom to keep pace with web technology as it undergoes its own evolution.

Being both flexible and sustainable, this cooperative approach enables the creation of a well-structured system for the organization of data, taking into account various cultural traditions and harnessing technological opportunities. Key figures in the sector may contribute alongside smaller organisations that are prepared to share their own experience and unique assets.

To comply with the new open philosophy of data sharing and reuse, the concept of traditional authority records must evolve as well. Previous discussion on whether authority control should be centrally or locally based, will morph as the focus shifts from a rigid conception of authority control to a more flexible notion of entity identification and relationships between entities. Organisations such as libraries, archives, museums, publishers and providers will engage with each other in the generation of new data and discovery of new resources, crossing the boundaries of specific domains to create data enrichment opportunities that would have been unthinkable before.

Likewise, the emerging BIBFRAME data model is currently the subject of much discussion and development within the library community and beyond. The intention behind the framework is to open up the possibilities of linked data to libraries, archives and museums, providing greater interoperability, visibility and discoverability for all types of resources held, taking into account the evolving guidelines and data models such as the IFLA Library Reference Model (LRM) and the Resource Description and Access (RDA) guidelines.

The aim of the components and the tools developed under the Share family of initiatives based on linked data, comprising Share-VDE (Virtual Discovery Environment), Share-ART (including the art history libraries of the Max Planck Institut), and Share-MUSIC (a pilot in the music domain), is to create an environment that is useful for the stewardship of research knowledge. Advanced discovery interfaces improve the user experience and deliver wider search results to library, archive and museum patrons, while librarians, archivists and museum curators will notice a number of advancements: from the continuing development of cataloguing functions native to Semantic Web standards to the integration of processes with local systems and tools for implementation in a collaborative environment. The goal is to help reveal the wealth of data within existing collections, which often remain hidden or unexpressed in a traditional catalogue.

Within the initiative, catalogue records of libraries, archives and museums are transformed to linked data through entity identification, reconciliation and data enrichment. Attributes are used to uniquely identify a Person, a Work or other Entities, with variations reconciled to form a cluster of data referring to the same Entity. These are then reconciled against traditional authority files and other external sources to enrich the records and create a network of information and resources. A database of relationships, open to the entire community, is compiled, as well as a common knowledge base of clusters that is accessible in RDF. The database uses the paradigms of the Semantic Web but also allows participating libraries to continue handling their own data as independently as possible.

Among the main areas of focus:

- enrichment of MARC record with URIs;
- conversion from MARC to RDF using the BIBFRAME vocabulary (and other additional ontologies as needed);

- creation of a virtual discovery platform with an adaptation of the BIBFRAME data model developed to provide a linked data discovery option;
- creation of a database of relationships and clusters of library materials (Sapientia Cluster Knowledge Base) accessible in RDF;
- implementation of tools for direct interaction with the data, permitting the validation, update, long-term control and maintenance of the clusters and of the URIs identifying the entities;
- batch/automated data updating procedures;
- batch/automated data dissemination to libraries;
- progressive implementation of further use cases in the priority order defined by the Share community, among these: copy cataloguing, original cataloguing, API for ILS's, retro conversion for local acquisition and administration systems based on MARC formats, and reporting.

The collaboration of participating institutions is encouraged so that decisions concerning subsequent steps for a full-scale rollout across library, archive and museum communities may be based on solid evidence from data sets. These data sets are not simply converted to RDF, but are also enriched with identifiers and interlinked, a prerequisite for use in the linked data environment. These processes can then contribute to building a realistic model of user activities, addressing any problems as well as the potential advantages of moving toward the linked data environment. Functionality dominates the design of the infrastructure, providing various environments and user interfaces for data creation, enrichment and supply workflows to the diverse group of participating librarians, professionals, scholars, researchers and students, encompassing a wide range of needs.

Casalini Libri (www.casalini.it) is a bibliographic agency, producing authority and bibliographic data; a library vendor, supplying books and journals, and offering a variety of collection development and technical services; and an e-content provider, working both for publishers and libraries.

@Cult (www.atcult.it) is a software development company, specializing in data conversion for linked data, and provider of Integrated Library System and Discovery tools, delivering effective and innovative technological solutions to improve information management and knowledge sharing.