Search and Analytics Challenges in Digital Libraries and Archives

VINCENZO MALTESE, FAUSTO GIUNCHIGLIA University of Trento, Italy

• Applied computing → Digital libraries and archives • Information systems → Decision support systems

Additional Key Words and Phrases: cataloging, data integration, knowledge graphs, analytics.

ACM Reference Format:

Vincenzo Maltese, and Fausto Giunchiglia, 2016. Search and Analytics Challenges in Digital Libraries and Archives. *ACM Journal of Data and Information Quality*.

1. MANAGING INFORMATION IN DIGITAL LIBRARIES AND ARCHIVES

Public institutions, such as universities, maintain data in several *information silos*, each of them engineered to serve a specific vertical application. Data about key entities - such as people, publications, courses, projects - is scattered across them and difficult to correlate due to the diversity in format, metadata, conventions and terminology used. In such scenario, nowadays it is practically impossible to correlate data and support *advanced search* and *analytics facilities*, in turn vital to identify institutional priorities and support institutional strategic goals, as well as to offer effective data visualization and navigation services to their users (e.g. researchers, students, alumni, companies).

A *catalogue*, in libraries and archives, is a collection of organized data describing the information content managed by an institution [Patton 2009]. Cataloging is the process (guided by rigorous rules) that information scientists follow to create and maintain metadata in order to effectively represent and exploit information content. The most widespread library data models are still traditional *record-based models*, i.e. models that bundle information about the same entity into a single record.

The advent of the Web opened boundless opportunities to information seekers, especially in terms of quantity of information and abundance of search tools. This has brought libraries and their cataloguing practices to a crisis point [Coyle and Hillmann 2007]. The enhanced users' expectations led them to embrace the Semantic Web vision [Berners-Lee et al. 2001]. It advocates that representing data in a uniform machine-readable format with explicit meaning allows the development of intelligent interconnected services, able to get and aggregate data from different sources. Libraries started adopting the Linked Data approach that in turn is leading to a paradigm shift from record-based to *entity-based models*, i.e. models in which relevant entities are assigned URIs and are described in terms of subject-property-object triples. All together triples form a *knowledge graph*. The extent to which this is happening is nicely described in [Alemu et al. 2012] and [Martin and Mundle 2014]. Active institutions include the British Library, the Deutsche Nationalbibliothek, the Oslo Public Library, and OCLC. WorldCat [Teets and Goldner 2013] constitutes a prominent example of such transition. The W3C continuously gives recommendations for libraries to get more involved in Linked Data projects. Library standards, such as RDA, FRBR, and BIBFRAME are now available in W3C standard formats (i.e. RDF).

2. CHALLENGES AND CRUCIAL REQUIREMENTS

The adoption of entity-based models poses new challenges that have a clear impact on the capacity to adequately manage data, and to offer innovative services to their users. Evidence of these difficulties can be found for instance in [Bygstad et al. 2009], [Byrne and Goddard 2010], [Alemu et al. 2012] and [Martin and Mundle 2014]. *Technological challenges* include the lack of tools and of a supporting platform. *Conceptual challenges* include the difficulty of identifying and of adopting the right

standards, e.g. in terms of data model and vocabularies. *Organizational challenges* mainly pertain to the obstacles that need to be overcome by institutions to move from consolidated practices and standards to new ones, the difficulty to allocate the necessary resources, to coordinate people with different skills and to face the lack of expertize. *Legal and security challenges* include the difficulty to comply with intellectual property rights (IPR), licensing and privacy obligations, as well as to trace data provenance and guarantee secure access to data. It has to be considered that there is always a trade-off between the degree of openness of a system and its capacity to offer useful services to users such that the right level of trust is kept.

We suggest that important crucial requirements to be met in order to adequately tackle these challenges, thus supporting institutions in the adoption of entity-based models and guarantee for adequate data and information quality, include:

- *Centralized access to information*. The traditional tendency of libraries to provide centralized access to information is currently disrupted by the data fragmentation [Barton and Mak 2012]. The catalogue should be designed to offer centralized access to information which is originally stored in different information silos and codified following different data models and formats. To populate the knowledge graph, the platform should provide data extraction, transformation and load (ETL), data correlation and merge facilities.
- *The definition of the data model*. The different institutional needs demand for the capacity to personalize the data model employed. To accommodate for all the key entities, the model should cover a broader range of entity types w.r.t. those traditionally maintained in catalogues [Giunchiglia et al. 2014]. This includes the capacity of the model to represent meta-information such as data provenance and users' preferences.
- *Authority control*. The system should support the well-established rules followed in libraries to control the form of names (name authority), establish identifiers (identity management) and standardize terminology (vocabulary control) [Patton 2009].
- *The development of a broader range of services*. Traditional discovery services offered by library catalogues, centered on the search for intellectual creations, are not sufficient to meet the enhanced user needs and expectations. In particular, there is a demand for systems offering advanced search [Giunchiglia et al. 2014], and analytics supporting institutional decision-making [Teets and Goldner 2013]. Interoperability services are needed to support the mapping with existing standards and the publication of Linked Data.

3. TOWARDS THE SOLUTION

In 2015 at the University of Trento we moved the first steps towards addressing this problem with the *Digital University* project. We are tackling the technological and conceptual challenges by developing a new platform able to support cataloguers in the definition of the data model, the authority control mechanisms, and (via APIs) the development of the services. This meets and extends the idea of *linking HUB* envisioned by [Byrne and Goddard 2010]. The system we developed provides an initial set of ETL, data correlation and merge facilities supporting the creation of a *knowledge graph* in adherence with the (locally defined) data model and authority control rules. Data about the same entity scattered across different datasets is correlated by means of identifiers or heuristics. Duplicates are detected and merged.

Though it comes by default with its own data model, the system supports the extension and personalization of the data model as function of the content types, the available data sources and the services that a given institution may want to offer. We are tackling the legal and security challenges by ensuring governance and privacy-by-design principles [Hoepman 2014] informed by the legal office of the University; dedicated data structures support provenance, users' preferences and permissions. Dedicated services will offer the capacity to select the data to be published as Linked Open Data, thus supporting interoperability and data reuse. We are tackling the organizational challenges by employing an interdisciplinary pool of people skilled in ICT and Library & Information Science that closely collaborate with representatives from the various departments. In particular, the latter provide the terminology that is used to develop the controlled vocabulary of each department. *Data scientists* [Davenport and Patil 2012] will be responsible of the metadata quality, the entity cataloging, as well as of the correct interpretation of data to support institutional decision-making. The services we are developing have the potential to improve the way the University knows itself, presents itself to the world, becomes more efficient and transparent.

REFERENCES

Alemu, G., Stevens, B., Ross, P., and Chandler, J. (2012). Linked Data for libraries: Benefits of a conceptual shift from library-specific record structures to RDF-based data models. New Library World, 113(11/12), 549-570.

Bygstad, B., Ghinea, G., and Klæboe, G. T. (2009). Organisational challenges of the Semantic Web in digital libraries: A Norwegian case study. Online Information Review, 33(5), 973-985.

Glenn E. Patton. 2009. Functional Requirements for Authority Data – A Conceptual Model. IFLA Working Group on Functional Requirements and Numbering of Authority Records. Walter de Gruyter.

Kristin E. Martin and Kavita Mundle. 2014. Positioning Libraries for a New Bibliographic Universe. Library Resources & Technical Services, 58(4), 233-249.

Tim Berners-Lee, James Hendler and Ora Lassila. 2001. The Semantic Web. Scientific American, 284(5), 28-37.

Karen Coyle and Diane Hillmann. 2007. Resource Description and Access (RDA): Cataloging rules for the 20th century. D-Lib magazine, 13 (1), 3.

Michael Teets and Matthew Goldner. 2013. Libraries' role in curating and exposing big data. Future Internet, 5 (3), 429-438.

Joshua Barton and Lucas Mak. 2012. Old Hopes, New Possibilities: Next-Generation Catalogues and the Centralization of Access. Library trends, 61(1), 83-106.

Fausto Giunchiglia, Biswanath Dutta, and Vincenzo Maltese. 2014. From Knowledge Organization to Knowledge Representation. Knowledge Organization, 41(1), 44-56.

Thomas H. Davenport and D. J. Patil. 2012. Data scientist: The sexiest job of the 21st century. Harvard Business Review, 90, 70-76.

Gillian Byrne and Lisa Goddart. 2010. The strongest link:Libraries and linked data. D-Lib magazine,16(11)

Hoepman Jaap-Henk 2014. Privacy design strategies. ICT systems security and privacy protection, Springer Berlin Heidelberg, 446-459.